UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Resilience: statistical study of psychosocial and biological predictors at the workplace

Heloísa Gabriela de Castro Vasconcelos Gonçalves Galante

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Lisete Sousa
Doutora Maria João Heitor

2019

Acknowledgments

Às minhas orientadoras, Professora Lisete e Doutora Maria João, pelo enorme apoio dado durante todo o processo deste projeto.

Aos meus pais, Luísa e Marc, e irmã Gabriela, que me acompanharam ao longo destes meses e sempre me compreenderam.

Aos meus avós, Fátima e Herculano, e Tia Nini, que me motivaram a ser forte e nunca desistir.

À minha Tia São, que não chegou a assistir ao final desta etapa, mas que decerto me vê.

Ao meu melhor amigo Bruno, que sempre soube como me sentia.

Ao meu companheiro de vida, Miguel, por tudo.

Resumo

A resiliência pode ser definida como a capacidade de um indivíduo adaptar-se com facilidade a infortúnios ou mudanças inesperadas. Também é vista como um mecanismo positivo de resposta a acontecimentos stressantes, desde desastres naturais a questões financeiras, problemas de saúde, divórcio e morte. A resiliência é estudada desde o século XIX e desde aí demonstra uma evolução conceptual, tendo sido concebida como trajetória, contínuo, sistema, característica, processo, ciclo e categoria qualitativa. Nas teorias que consideram a resiliência uma característica, acredita-se que uma combinação de outras características físicas e psicológicas potenciem a capacidade de ser resiliente. Perante a cultura do trabalho que se verifica no mundo contemporâneo, é interessante estudar o papel da resiliência em trabalhadores, cujo maior fator de stress é o trabalho em si e todas as suas componentes e ligações com as restantes áreas da vida pessoal, de modo a entender de que forma estes diversos fatores podem influenciar os seus níveis de resiliência.

O principal objetivo deste estudo é identificar possíveis preditores de resiliência em trabalhadores. Com este intuito, os 1385 colaboradores de uma instituição bancária que cumpriam os critérios de inclusão pré-definidos foram convidados a participar no projeto. Aos 260 que aceitaram fazer parte do estudo e realizar exames laboratoriais, foi aplicado um questionário pormenorizado acerca do respondente, do seu trabalho, experiências e vida pessoal, incluindo diversos tipos de variáveis (sociodemográficas, relacionadas com o trabalho, relacionadas com o estilo de vida, clínicas), para além de diversas escalas validadas referentes a estes temas (ASSET, MHI-5, Escala de Satisfação no Trabalho, CAGE, Escala de Felicidade Subjetiva, OSLO, Presentismo e Absentismo). A isto, juntam-se as variáveis bioquímicas provenientes dos resultados das análises laboratoriais. A variável resiliência foi medida através das versões de 25 e 10 itens da Escala de Resiliência de Connor-Davidson. Estes dados foram recolhidos entre Novembro de 2012 e Junho de 2013. A base de dados contendo todas as variáveis foi corrigida e validada no contexto deste projeto antes de se iniciar o processo de análise abaixo descrito.

Primeiramente, foi realizada uma análise exploratória dos dados, de modo a caracterizar os respondentes. Seguidamente, decidiu-se qual das versões da escala de resiliência seria considerada a medida única de resiliência durante o decorrer da análise, através do cálculo do coeficiente de correlação ordinal de Spearman e do respetivo teste de significância. Os resultados demonstram que o valor desta medida sugere uma associação elevada entre ambas as versões, o que significa que a menos redundante, de 10 itens, deve ser escolhida. Depois, devido ao grande número de possíveis preditores na base de dados, procedeu-se a uma pré-seleção das variáveis cuja associação com a resiliência (se numéricas, usando o coeficiente de correlação de Spearman) ou diferença entre grupos relativamente à resiliência (se categóricas, usando os testes de Mann-Whitney ou Kruskal-Wallis) fossem estatisticamente significativas. As variáveis pré-selecionadas foram integradas num modelo de regressão inicial, ao qual se aplicou a seleção stepwise de modo a obter-se um modelo mais parcimonioso que simultaneamente explicasse a maior percentagem possível da variabilidade da resiliência. Este modelo final, validado no que diz respeito aos pressupostos de uma análise de regressão, engloba como preditores a saúde mental, segurança no trabalho e sobrecarga laboral (medidos por subescalas do ASSET), ter interesses ou hobbies, tomar medicação para a ansiedade crónica e o nível de presentismo, e explica aproximadamente 35% da variabilidade da resiliência. Possíveis interações a adicionar a este modelo final foram analisadas relativamente aos preditores descritos, mas não se demonstraram estatisticamente significativas na análise de regressão. Por último, com o objetivo de entender de melhor forma a estrutura das variáveis

selecionadas inicialmente, foi aplicada uma análise fatorial múltipla de dados mistos. Esta metodologia inovadora propõe alargar o conceito da análise fatorial múltipla, em que existem vários grupos de variáveis mas em que cada grupo contém apenas variáveis do mesmo tipo, para os casos em que estes grupos acomodam tanto variáveis do tipo numérico como categórico. Os grupos formados com as variáveis em questão foram “Trabalho” e “Saúde Física e Mental”. As três primeiras componentes principais explicam pouco mais de 20% da variabilidade dos dados. Relativamente à primeira componente principal, o impacto de ambos os grupos é quase idêntico. O grupo "Trabalho" contribui mais para a segunda componente principal, enquanto que para a terceira o grupo "Saúde Física e Mental" apresenta a maior contribuição, embora a diferença seja bastante pequena.

As limitações deste estudo prendem-se essencialmente com a pequena dimensão da amostra, com o facto de não ter sido recolhida aleatoriamente e por ser proveniente de um tipo muito limitado de trabalhadores, o que pode enviesar os resultados. Não obstante, a análise realizada foi capaz de dar resposta aos objetivos do projeto e revela-se importante e relevante para um maior conhecimento do fenómeno da resiliência e da sua importância nos trabalhadores. Futuros estudos com amostras de maior dimensão e mais variabilidade de trabalhadores podem consolidar e confirmar os resultados obtidos, aprofundando esta temática.

Palavras-chave: resiliência, dados categóricos, métodos não-paramétricos, regressão múltipla, análise fatorial múltipla de dados mistos.

Abstract

Resilience is one of the most important characteristics of an employee in the workaholic culture of our current society. The main purpose of this study is to discover which other traits, habits and features can significantly influence and impact resilience levels. For this purpose, a comprehensive questionnaire was applied to 260 workers from a banking institution between November 2012 and June 2013, including sociodemographic, work-related, lifestyle-related, clinical and biochemical variables, while also comprising several validated scales. Resilience was measured with the 25-item and 10-item versions of the Connor-Davidson Resilience Scale. After a pre-selection of the survey's variables with a statistically significant association (if numerical) or difference between groups (if categorical) regarding resilience, its best predictors were identified through a regression analysis: ASSET's Psychological wellbeing, Job security and Overload scales, having interests/hobbies, taking medication for chronic anxiety and the percentage of work performance loss (presenteeism). This regression model explains about 35% of resilience's variability. Also, in an attempt to understand the structure of the resilience predictors and reduce its dimension, a multiple factor analysis of mixed data was conducted regarding the pre-selected variables, which were divided in two conceptual groups: "Work" and "Physical and Mental Health". The first three principal components explain about 20% of their variability. This study was important to provide more evidence and information regarding resilience predictors at the workplace and exploring the relationships between resilience and several scales, some of which have not been analyzed by the scientific community so far. However, further studies with larger sample sizes, mixed categories of workers and other types of variables are needed to confirm the obtained results. This knowledge can lead to improvements in workers' resilience levels, and therefore increase productivity and work satisfaction of a company's employees, which is fruitful to both.

Keywords: resilience, categorical data, non-parametric methods, multiple regression, multiple factor analysis of mixed data.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 Definition

Resilience can be defined as the capability to adapt with ease to misfortune or unexpected change (Merriam-Webster, 2019). It is a coping mechanism or response to stressful experiences, that can be related to any type of problem or challenge, ranging from natural disasters to financial issues, health concerns, divorce and death.

The concept of resilience has been studied since the 1800s and is continuously evolving. During its development, resilience has been constructed as a trajectory, a continuum, a system, a trait, a process, a cycle and a qualitative category (Jackson et al., 2007). In theories that consider resilience a trait, it is believed that a combination of physical and psychological characteristics, including body chemistry and personality factors, gives individuals the skills to be resilient (Jacelon, 1997).

Resilient people tend to not become as bothered, upset or fearful when adversities happen as people with less resilience do, although that does not mean they do not feel them as deeply. Having resilience merely implies that the subject of a certain unpleasant event utilizes their skills, strengths and knowledge to overcome it and grow from it, without losing hope or falling into despair. Figure 1.1.1 provides an illustration of the most common characteristics and states conceptually associated to resilience.

Figure 1.1.1: Resilience model (George Mason University's Resilience Model, n.d.).

1.2 Literature review

Resilience at the workplace

According to the World Health Organization (1994), the majority of the world's population (58%) spends one-third of their adult life at work. This allows the achievement of material and economic goals and provides a better quality of life. However, there are several jobs in many countries that still entail hazards to health, therefore reducing the well-being, working capacity and life span of working individuals.

In a workplace setting, various demanding situations can take place: heavy workload, impractical deadlines, poor communication, rigid schedules, competition between colleagues, discrimination and general bad work environment. These can lead to workers' discouragement, lack of motivation and mental and physical health problems. To deal with this constant pressure and stress, resilience is a needed and valuable skill. It has been associated with various positive states, including optimism, zest, curiosity, energy and openness to experience (Tugade & Fredrickson, 2004), which can boost performance and creativity.

Work-related psychosocial risk factors or stressors are likely to have an impact in physical and mental health through closely interrelated emotional, cognitive, behavioral and physiological mechanisms. Strong evidence was found that high job demands, low job control, low co-worker support, low supervisor support, low procedural and relational justice and a high effort–reward imbalance predict the incidence of stress-related mental disorders, for which resilience can be a structural protective factor (Rutter, 2006).

Environmental, demographic and lifestyle factors

Tugade and Fredrickson (2004) suggest that the ability to find positive meaning in adverse situations and to regulate negative emotions contributes to personal resilience. Studies have shown a link between low resilience and several mental health outcomes, such as burnout, secondary traumatic stress, depression and anxiety (Rees et al., 2015). Furthermore, evidence on resilient survivors of violent trauma shows that these exhibited better health and less severe post-traumatic stress disorder symptoms than those who were not resilient (Connor, Davidson & Lee, 2003).

Characteristics like sex, education level, income level and history of childhood abuse are thought to contribute to the prediction of resilience: females, individuals with lower levels of education and income and individuals with history of childhood trauma report diminished resilience (Sparks et al., 1997). Other factors such as age, race/ethnicity, substance use, social support, chronic diseases, recent life stressors and past traumatic events are also shown to impact individuals' level of resilience (Bonanno et al, 2007). However, some studies claim there is no relationship between resilience and social support and lifestyle or work-related factors (Corina & Adriana, 2013; Black et al., 2017), although it is generally acknowledged that resilience moderates various stress types (Liu et al., 2018).

As far as individual personality traits go, a negative correlation between resilience and neuroticism and positive associations with extraversion and conscientiousness have been accounted for (Fayombo, 2010). Bonanno et al. (2002) discusses hardiness, self-enhancement, repressive coping, positive emotions and laughter as being resilience-promoting.

1.3 Research objectives

The general objective of this study is the identification of sociodemographic, psychological, clinical and biological factors that impact workers' resilience and global mental health. This study aims to answer the following questions:

1. What factors impact the resilience level at the workplace?
2. Is there an underlying structure to the main resilience predictors?

1.4 Project outline

This project begins with the introduction above, stating the study's subject, the state of the art and reasons for its importance and relevance in today's society, followed by a thorough description of the dataset. Next, a methodology chapter describes the theoretical principals behind the statistical methods used. This is followed by the results chapter, where the outcomes of the statistical analyses are presented and compared. It ends with a discussion chapter, where the obtained results are critically examined and further work and research regarding this subject are suggested.

1.5 Study Design and Description

The study presented in this project is descriptive, with a cross sectional design. It was conducted in the context of the project "Health Impact Assessment of Employment Strategies", led by Doctor Maria João Heitor dos Santos in 2011, resulting of a collaboration protocol between the then High Commissioner of Health, the Institute of Preventive Medicine and Public Health of the Faculty of Medicine of the University of Lisbon and the National Institute of Health Doutor Ricardo Jorge. It contains data from a survey applied to employees of Caixa Económica Montepio Geral (CEMG) in the Lisbon County, between November 2012 and June 2013. The questionnaire itself is not included in the Appendices section due to it being confidential.

1.5.1 Sample

Eligible subjects for the study were identified from the employee population of CEMG using the following selection criteria:

- Employees between the age of 18 and 69 years;
- Employees able to understand and sign an informed consent.

A total of 1385 Lisbon County's CEMG employees were identified in these conditions. All were sent an invitation to participate in the survey through their institutional email, and only those who voluntarily wanted to take part in the project were included in the study. Therefore, the resulting sample is non-random.

From the 1385 invited employees, 405 responded to the survey. A blood sample and anthropometric measures were also collected from 260 participants out of the 405. Matching of participants between the survey and blood samples was possible using an ID linked to a name and email address, to guarantee confidentiality.

In order to include the variables collected from the blood sample and biochemical parameters in this study, only the data from the 260 participants was used in the analysis.

1.5.2 Description of the variables

From the original dataset, containing 406 raw variables, a subset of 106 variables indicated by the project leader was considered for this analysis. This set contains variables which are reported by experts and literature as being of interest regarding the study of the resilience phenomenon.

The final dataset contains the following types of variables (detailed in Appendix A):

- **Sociodemographic**
- **Work-related**
- **Lifestyle-related**
- **Clinical**
- **Biochemical**
- **Scales**

- **CDRISC Scale** (Connor & Davidson, 2003), a brief self-rated assessment to quantify individual resilience. The CDRISC is composed of 25 items, each rated on a 5-point scale of responses (0–4). The scale is rated based on how the subject felt over the previous month. The total score ranges from 0–100, with higher scores reflecting greater resilience;
- **ASSET** (Cooper, Sloan & Williams, 1988), a short stress evaluation tool to assess the risk of workplace stress, containing twelve subscales: eight of which evaluate the workers' job perception, two of which evaluate their attitude towards the organization and the other two which evaluate their physical and mental health;
- **Mental Health Index Scale - MHI-5 (**Veit & Ware, 1983; Ribeiro, 2001), a scale of 5 items, each rated in a 6-point scale, used for the measurement of mental health status;
- **Oslo Social Support Scale** (Dalgard, 1996; Dalgard et al. 2006), a scale of 3 items, two of them rated in a 5-point scale and one rated in a 4-point scale, that allows overall assessment of social support;
- **Subjective Happiness Scale** (Lyubomirsky & Lepper, 1999), a scale of 4 items, each rated in a 7-point scale, which evaluates one's self-assessment of subjective happiness;
- **CAGE Scale** (Ewing, 1984), a scale of 4 items, each rated as 0 (if the answer is "No") or 1 (if the answer is "Yes"), to measure alcohol consumption;
- **Job Satisfaction Scale** (Cooper, Sloan & Williams, 1988), a scale of 7 items, each rated in a 7-point scale, that evaluates one's level of satisfaction with their job;
- **Presenteeism Scale** (Kessler et al. 2004; 2003), a scale of 3 items, each ranging from 0 to 10, that assesses the low quality of work due to poor health or mental health status.

1.5.3 Data preparation

- **Database cleaning**

  Prior to the analysis itself, an internal data cleaning was conducted, in order to identify inaccurate or incorrect records and proceed to their correction or removal, to assure the consistency of the database.

- **New variables**

  Based on the objectives of this project, a set of extra indicators was computed from some of the original variables in the dataset:

- **CDRISC-10:** A new variable for resilience was created, using the 10 item version of the CDRISC. This version is obtained from the 25 item, based on a psychometric analysis that allowed the identification of the 10 items that best captured the features of resilience with minimal redundancy. The 10 item version (final scores range from 0 to 40) comprises items 1, 4, 6, 7, 8, 11, 14, 16, 17, 19 from the original scale.
- **Absenteeism:** Absenteeism refers to the usual pattern of absence in a function or obligation (Weiner, Schmitt & Highhouse, 2012). To create an absenteeism indicator, the difference between the number of effective work hours and the number of expected work hours was computed. Thus, negative values represent loss in work hours, positive values represent surpluses in work hours and 0 represents absence of loss (or surplus) of work hours.

- **Cardiac Risk Index**: This indicator is based on the standards of the Portuguese Society of Cardiology, according to which its risk factors are:

  Modifiable factors:

  Smoking;

  High blood pressure (pre-hypertensive and hypertensive);

  Diabetes;

  Obesity;

  High cholesterol;

  Sedentarism;

  Psychosocial stress.

  Non-modifiable factors:

  Male gender;

  Age greater than 55 years;

  Heredity.

For each risk factor present in an individual a value of 1 was considered, except for heredity because no information was available. A cardiac risk indicator was computed through the sum of the values for the set of risk factors. Therefore, the values of this indicator range from 0 to 9, and the higher the value, the greater the cardiovascular risk.

Besides the variables above, there was also the need to work with the total score for each scale comprised in the survey, instead of the individual questions' scores. For this purpose, the following scale scores were calculated:

- **ASSET subscales:** First, each of the subscales' sten score (Colman, 2019) was computed using the individual level norms for the general population group, according to the ASSET Norm Supplement. Then, to obtain final scores, the following rules were applied to the "perceptions of your job", "attitudes to your organization" and "your health" scales, respectively:

  Scores below sten 3 indicate very low levels of the stressor/ very low levels of commitment/ very good health levels (coded as 1);
  Scores below sten 4 indicate low levels of the stressor/ low levels of commitment/ good health levels (coded as 2);
  Scores within the range defined by sten 4 to sten 7 indicate average levels of the stressor, commitment and health (coded as 3);
  Scores above sten 7 indicate high levels of the stressor/ high levels of commitment/ poor health levels (coded as 4);
  Scores above the sten 8 indicate very high levels of the stressor/ very high levels of commitment/ very poor health levels (coded as 5).

- **MHI-5**: The questions "How much of the time in the previous 4 weeks have you felt calm and peaceful?" and "How much of the time in the previous 4 weeks have you been a happy person?" are reversed (a 6 turned into a 1, a 5 into a 2, a 4 into a 3, a 3 into a 4, a 2 into a 5 and a 1 into a 6). Then, the mean of the 5 items is multiplied by 100 and divided by 5, varying from 0-100. Individuals with a score lower or equal to 52 present psychological distress;
- **OSLO-3:** A sum index was computed using the raw individual scores of the 3 questions, ranging from 3 to 14. A score of 3-8 means "poor support", 9-11 means "moderate support" and 12-14 means "strong support";
- **Subjective Happiness Scale:** To score the scale, the question "Some people are generally not very happy. Although they are not depressed, they never seem as happy as they might be. To what extent does this characterization describe you?" was reversed (a 7 turned into a 1, a 6 into a 2, a 5 into a 3, a 3 into a 5, a 2 into a 6 and a 1 into a 7), and the mean of the 4 items was calculated;
- **CAGE**: The final score corresponds to the sum of the 4 items. A sum greater than or equal to 2 indicates a high probability of alcohol dependence;
- **Job Satisfaction Scale:** An overall indicator was obtained by calculating the arithmetic mean of the seven items;
- **Presenteeism Scale:** Absolute Presentism corresponds to the proportion of the loss of one's performance at work, calculated by multiplying by 10 the response to the question "On a 0-to-10 scale, how would you rate your overall job performance on the days you worked during the

past 4 weeks (28 days)?". This indicator varies between 0 and 100, and the higher its value, the lower the performance loss.

1.5.4 Software

The database preparation and statistical analysis were conducted using Microsoft Excel and R Studio (R Core Team, 2019), respectively.

1.5.5 Ethical considerations

This study was approved by two institutional ethical committees: the Ethics Committee for Health of the National Institute of Health Doutor Ricardo Jorge (INSA) and the Ethics Committee for Health of the Lisbon/North Hospital Center of the Faculty of Medicine of the University of Lisbon (CHLN/FMUL). It was also approved by the National Commission of Data Protection (CNPD). All participants signed an informed consent.

1.5.6 Previous studies in the context of the project

The data corresponding to the 405 questionnaire respondents has been initially studied, in the context of the original project. The CDRISC scale was validated for a Portuguese sample (Faria Anjos et al., 2019), where the three identified dimensions of resilience were the self-efficacy, spirituality and social support factors present in the scale. The Portuguese version of ASSET was validated in terms of its psychometric properties and its convergent validity was tested (Heitor et al., 2018).

Chapter 2

Statistical Methodology

This project is based on workers' responses to a comprehensive questionnaire regarding several dimensions of their life. Survey variables are mostly categorical, which can be limiting as far as statistical methodology goes – most studies with a similar background stick to scale validations and contingency table analyses. In this study, the methods were chosen taking into account its research questions and the quality of the data.

Firstly, as there are two measures for resilience, the Spearman rank correlation coefficient was calculated with the purpose of selecting which should be considered the most adequate dependent variable for a future regression model. This method was also used to select the relevant numerical independent variables regarding resilience and to check for potential associations between them. Concerning the categorical independent variables, Mann-Whitney (for variables with only two groups) and Kruskal-Wallis (for variables with more than two groups) tests were applied to check for significant differences in resilience between groups. For the Kruskal-Wallis tests with a significant p-value, Conover's test was used in order to identify which of the respective variables' groups present statistically significant differences in resilience levels. Throughout these analyses, the Benjamini-Hochberg correction was applied to correct p-values in cases of multiple testing. A multiple regression analysis was conducted to study the linear relationship between resilience and the selected predictor candidates. To check for plausible interaction terms to add to this regression model, Fisher's exact test was applied to determine if there was a statistically significant association between any two pairs of categorical variables present in the final multiple regression model. The methodologies described so far mean to answer the first research objective formulated in Section **1.3**. Finally, a multiple factor analysis of mixed data was conducted to reduce dimensionality and explore the structure of the initially selected variables, considering the presence of both categorical and numerical variables and their potential to form conceptual groups. This method aims to answer the second research objective mentioned in Section **1.3**.

The R functions and respective packages that allowed the implementation of these techniques are mentioned throughout the text. The significance level $\alpha = 0.05$ was used for all methods. Also, it should be mentioned that the Levene and Shapiro-Wilk tests for the validation of the ANOVA and linear regression assumptions are not detailed in this chapter because they are widely known and not the main focus of the present statistical analysis.

2.1 Spearman rank correlation coefficient

The Spearman rank correlation coefficient (r_s) represents a non-parametric alternative to the Pearson product-moment correlation. This measure can be applied to numerical or ordinal variables and is robust when extreme values are present. If there is an intention to perform hypotheses tests over the population's correlation coefficient, Spearman's coefficient can also substitute Pearson's coefficient

when the latter's assumptions of bivariate normality or linearity are not verified by the data (Hauke & Kossowski, 2011).

Spearman's rank correlation coefficient assesses how well an arbitrary monotonic function can describe a relationship between two variables without making assumptions about their probability distributions. It can range from -1 to +1, where a value of 0 indicates that there is no linear association between ranks, and, therefore, no monotonic relationship between the variables, a value of -1 indicates a perfect negative correlation between ranks and a value of +1 represents a perfect positive correlation between ranks (Table 2.1.1).

Table 2.1.1: Interpretation of Spearman's coefficient according to Fowler, Cohen and Jarvis (2009).

$\lvert r_s \rvert$	**Association strength**
0.90 – 1.00	Perfect
0.70 – 0.89	Strong
0.40 – 0.69	Moderate
0.20 – 0.39	Weak
0.00 – 0.19	Very weak

Let there be X and Y, a pair of random variables. For a sample of size n, the n raw scores x_i, y_i, $i = 1, \dots, n$, are converted to ranks. The formula for r_s when there are no tied ranks is:

$$r_s = 1 - \frac{6 \sum {d_i}^2}{n(n^2-1)} \qquad (2.\ 1)$$

where r_{x_i} is the rank of x_i, r_{y_i} is the rank of y_i and $d_i = r_{x_i} - r_{y_i}$.

The formula to use when there are tied ranks is:

$$r_s = \frac{\sum_i (r_{x_i} - \bar{r_x})(r_{y_i} - \bar{r_y})}{\sqrt{\sum_i (r_{x_i} - \bar{r_x})^2 (r_{y_i} - \bar{r_y})^2}} \qquad (2.\ 2)$$

To test whether the Spearman correlation coefficient is significantly different from zero, the following is hypothesized:

H_0: There is no association between X and Y;

H_1: There is an association between X and Y.

For a large number of (X_i, Y_i) pairs, the distribution of $T = R_s \sqrt{\frac{n-2}{1-{R_s}^2}}$ can be approximated by a Student's t distribution with $n - 2$ degrees of freedom, under the null hypothesis. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $|T_0| \geq t_{(n-2,\ 1-\alpha/2)}$, which represents the $1 - \alpha/2$ quantile of a Student's t distribution with $n - 2$ degrees of freedom.

The function used in R was *cor.test()* from the stats package.

2.2 One-way analysis of variance

The one-way analysis of variance, or one-way ANOVA, is a parametric method that allows the comparison of the population means of three or more independent groups, in order to determine whether there is statistical evidence that these are significantly different. This test presumes the existence of a continuous dependent variable and a categorical independent variable with k mutually exclusive levels, $k = 3,4, \ldots$, of sizes n_i, $i = 1, \ldots, k$, where $\sum n_i = N$.

One-way ANOVA assumptions are as follows:

- Independence of observations;
- Random sample of data from the population;
- Normal distribution of the dependent variable across factor groups;
- Homogeneity of variances across factor groups;
- No outliers.

These assumptions are explained in more depth in Section **2.5** of this chapter.

The null and alternative hypotheses of a one-way ANOVA can be expressed as:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$;

H_1: At least one μ_i differs from the remaining.

where μ_i is the population mean of the ith group ($i = 1,2, \ldots, k$).

The test statistic for a one-way ANOVA is denoted as F. For an independent variable with k groups, the F statistic evaluates whether the group means are significantly different. Its components are usually depicted in a table like the following:

Table 2.2.1: ANOVA components.

	Sum of Squares (SS)	**Degrees of freedom**	**Mean Square (MS)**	**F**
Group	$\sum_{j=1}^{k} n_j(\bar{y}_{.j} - \bar{y}_{..})^2$	$k - 1$	$SS_{Treat}/(k-1)$	$MS_{Treat}\ /MS_{Error}$
Error	$\sum_{j=1}^{k}\sum_{i=1}^{n_i}(y_{ij} - \bar{y}_{.j})^2$	$N - k$	$SS_{Error}/(N-k)$	
Total	$SS_{Treat} + SS_{Error}$	$N - 1$		

where y_{ij} is the ith observation of the jth group, $\bar{y}_{.j}$ is the mean of the jth group and $\bar{y}_{..}$ is the overall mean of the N observations.

The F statistic, calculated by $F = MS_{Treat}\ /MS_{Error}$, follows a F-Snedecor distribution with $k - 1, N - k$ degrees of freedom, under the null hypothesis. For a significance level of α, the null hypothesis

is rejected if the observed value of the test statistic $F_0 \geq F_{(k-1,\ N-k,\ 1-\alpha)}$, which represents the $1-\alpha$ quantile of a F-Snedecor distribution with $k-1, N-k$ degrees of freedom.

If the test p-value is significant, sample contrasts or post-hoc tests can be used in order to determine which of the sample pairs are significantly different. However, the Type I error rate tends to become inflated when performing these methods, which raises concerns about multiple comparisons.

The function used in R was *anova()* from the stats package. The homogeneity and normality assumptions were tested by running *leveneTest()* and *shapiro.test()*, from car and stats packages, respectively.

2.3 Mann-Whitney U test

The Mann-Whitney U test is a non-parametric alternative to the independent sample t-test, used when its assumptions are not met or when data are ordinal. This test can be applied when measuring the same dependent variable in two independent populations (X and Y) to assess if there are differences between them (Mann & Whitney, 1947). Therefore, the test's hypotheses are:

H_0: X and Y have the same distribution;

H_1: The distributions of X and Y differ on location.

This test assumes the following criteria:

- Random samples from the populations;
- Independence within samples and mutual independence between samples;
- Measurement scale of the dependent variable is at least ordinal.

The procedure to compute the observed value of the test statistic is:

1. All the observations are ranked, beginning with 1 for the smallest value. When there are groups of tied values, a rank equal to the midpoint of unadjusted rankings is attributed to each group;
2. The sum of the ranks for the observations from sample 1 is calculated. The sum of ranks in sample 2 is now determined, since the sum of all the ranks equals $N(N+1)/2$, where N is the total number of observations;
3. The U statistic is given by $\min(U_1, U_2)$, where U_i is given by:

$$U_i = R_i - \frac{n_i(n_i+1)}{2} \qquad (2.3)$$

where n_i is the size of the ith sample and R_i is its rank sum ($i = 1,2$).

For large samples, the distribution of the test statistic $Z = \frac{U-\mu_U}{\sigma_U}$, where $\mu_U = \frac{n_1 n_2}{2}$ and $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}$, is approximated by a Standard Normal distribution, under the null hypothesis. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $|Z_0| \geq z_{1-\alpha/2}$, which represents the $1-\alpha/2$ quantile of a Standard Normal distribution.

The function used in R was *wilcox.test()* from the stats package.

2.4 Kruskal Wallis test

The Kruskal-Wallis test is a non-parametric alternative to a one-way ANOVA. It assesses whether a number of populations originate from the same distribution. It is used to compare three or more independent populations based on samples with equal or different sample sizes, therefore representing an extension of the Mann–Whitney U test (Kruskal & Wallis, 1952).

For $k = 3, 4, \ldots$ samples of size $n_i, i = 1, \ldots, k, n = \sum n_i$:

H_0: All k populations have the same distribution;

H_1: At least two of the k populations differ in location.

1. Data is ranked from 1 to n ignoring group membership. Tied values are assigned the average of the ranks they would have received had they not been tied.
2. If the data contain no ties, the test statistic is given by:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(n+1) \qquad (2.4)$$

In the presence of ties, a corrected test statistic can be applied (Siegal and Castellan, 1988):

$$H = \left(\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{{R_i}^2}{n_i} - 3(n+1) \right) \Big/ \left(1 - \frac{\sum_{j=1}^{g} ({t_j}^3 - t_j)}{n^3 - n} \right) \qquad (2.5)$$

where R_i is the rank sum of the ith group, g is the number of distinct groups of ties and t_j is the number of ties in the jth group, $j = 1, \ldots, g$.

In case of no ties, the observed value of H should be compared to the critical value obtained from the exact distribution of H. Otherwise, this distribution can be approximated by a Chi-squared distribution with $g - 1$ degrees of freedom under the null hypothesis, although the approximation is unsatisfactory when n_i values are small. In this case, for a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $H_0 \geq {\chi^2}_{(g-1,1-\alpha)}$, which represents the $1 - \alpha$ quantile of a Chi-squared distribution with $g - 1$ degrees of freedom.

If the test p-value is significant, then at least two of the k populations differ in location. In these cases, post-hoc tests can be used in order to determine which of the sample pairs are significantly different. However, the Type I error rate tends to become inflated when performing these methods, which raises concerns about multiple comparisons.

The function used in R was *kruskal.test()* from the stats package.

2.5 Conover's post-hoc test

The Conover test is a non-parametric post-hoc test for multiple comparisons. Meant to follow a Kruskal-Wallis test when the null hypothesis is rejected, it determines which groups differ significantly. It is also statistically more powerful than other non-parametric alternatives, such as Dunn's test (Conover & Iman, 1979).

For every pair of groups $(i,j), i \neq j,\ i,j = 1,\dots,k$ of a categorical variable with k groups, this test hypothesizes:

H_0: $\mu_i = \mu_j$;

H_1: $\mu_i \neq \mu_j$.

with a total of $k(k-1)/2$ possible hypotheses.

This test uses the following test statistic:

$$T = \frac{|\bar{R}_i - \bar{R}_j|}{s.e.} \tag{2. 6}$$

where $s.e. = \sqrt{\frac{1}{n-1}\left[\sum {R_i}^2 - n\frac{(n+1)^2}{4}\right]\frac{n-1-H}{n-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$.

Here, n represents the total sample size, R_i is the rank sum of the ith group, n_i and n_j are the sizes of the groups being compared, $\bar{R}_i$ and $\bar{R}_j$ are their respective mean ranks and H is the test statistic from the Kruskal-Wallis test (with or without ties). Under the null hypothesis, the distribution of T is approximated to a Student's t distribution with $n-k$ degrees of freedom. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $T_0 \geq t_{(n-k,1-\alpha/2)}$, which represents the $1-\alpha/2$ quantile of a Student's t distribution with $n-k$ degrees of freedom.

The function used in R was *kwAllPairsConoverTest()* from the PMCMRplus package.

2.6 Benjamini-Hochberg correction

The Benjamini-Hochberg correction is one of many methods designed to correct p-values after multiple statistical tests. This correction in particular is based on the False Discovery Rate (FDR), which is defined as the expected proportion of falsely rejected hypotheses among the set of rejected hypotheses if there is at least one rejection, and zero otherwise (Benjamini & Hochberg, 1995).

To apply the BH procedure, we first test each of the $m = k(k-1)/2$ hypotheses under consideration by calculating a test statistic and comparing it to the appropriate distribution to obtain a p-value. Let $p_{(i)}, i = 1,\dots,m$ be the ordered p-values, and $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. Then, to obtain a FDR control level α^*, we reject all $H_{(i)}$ for $i = 1,\dots,k$, for which:

$$k = max\left\{i: p_{(i)} \leq \frac{i}{m}\alpha^*\right\} \tag{2. 7}$$

and reject no hypotheses if this maximum does not exist. This procedure controls the FDR at α for any configuration of false null hypotheses, assuming independent test statistics.

In R, the correction was applied selecting *p.adjust="BH"* within the functions whose tests include multiple comparisons.

2.7 Multiple Linear Regression

Multiple Linear Regression is a statistical procedure that uses several explanatory variables to predict the outcome of a response variable. The goal of this technique is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. It represents an extension of the simple linear regression, which only contains one explanatory variable.

Given a response variable Y with $i = 1, \dots, n$ observations and p explanatory variables $\{X_1, X_2, \dots, X_p\}$, the formula for a multiple linear regression model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad (2.\ 8)$$

where y_i is the ith observed value of the response variable, x_{ik} is the ith value of the X_k explanatory variable ($k = 1, \dots, p$), $\{\beta_0, \beta_1, \beta_2, \dots, \beta_p\}$ are unknown parameters (β_0 is the intercept and the remaining are the respective coefficients of the X_k, $k = 1, \dots, p$, explanatory variables) and ε_i are the random errors. This equation can also be expressed in a matrix form:

$$\mathbf{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2.\ 9)$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is the $n * 1$ vector of the response variable, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ is the $(p + 1) * 1$ vector of the regression parameters, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ is the $n * 1$ vector of the random errors and

$\mathrm{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}$ the $n * (p + 1)$, $(n \geq p)$ is the design or regression matrix.

The expressions $\widehat{\boldsymbol{\beta}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathbf{Y}$, $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathrm{X}\widehat{\boldsymbol{\beta}}$ and $\mathrm{S}^2 = \hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathrm{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{Y} - \mathrm{X}\widehat{\boldsymbol{\beta}})}{n-p-1}$ represent the estimated $\{\beta_1, \beta_2, \dots, \beta_p\}$, $\boldsymbol{\varepsilon}$ (also called residuals) and variance of the model, respectively, obtained by applying the least squares method.

A multiple linear regression analysis assumes:

- Normality of errors

  This assumption can be validated with the Shapiro-Wilk test (Shapiro & Wilk, 1965), which tests the null hypothesis that a sample came from a Normally distributed population. Monte Carlo simulation was used to show that Shapiro–Wilk has the best power for a given significance level when compared to Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests (Razali & Wah, 2011).
  A graphical way to analyze this assumption is to obtain a boxplot of the residuals or a quantile-quantile plot, that represents the theoretical quantiles of a Normal distribution *vs* the model's residuals.

- Homoscedasticity of errors

  This assumption can be validated with Levene's test (Levene, 1960), which tests the null hypothesis that, for a variable calculated for two or more groups, the population variances are equal. This test is robust and mostly well accepted by the statistical community, although it is not asymptotically distribution free (O'Brien, 1981).
  A graphical way to analyze this assumption is to plot predicted values *vs* residuals and residuals *vs* independent variables. A plot whose observations resemble a wedge-shape suggest heteroscedasticity.

- Independence of errors

  Errors should be independent, meaning they are not capturing some information about the model. If this is not true, it will lead to an inaccurate model. Graphically, this assumption can be validated by observing the plot of observation indices *vs* observed values of the dependent variable.

- Linear relationship between the dependent variable and the independent variables;

  Nonlinearity is usually most evident in a plot of observed values of the independent variable *vs* observed values of the dependent variable.

- Absence of multicollinearity (high correlation between independent variables).

  One of the ways to validate this assumption is to calculate the Variance Inflated Factors (VIFs). For the kth predictor, $VIF_k = \frac{1}{1-{R_k}^2}$, where ${R_k}^2$ is the R^2 value obtained by running a regression of the kth predictor over the remaining predictors. A VIF of 1 means there is no correlation between the kth predictor and the other independent variables, while VIFs higher than 4 should be investigated.

To run the regression analysis, the function used in R was *lm()* from the stats package.

Table 2.7.1: ANOVA components for multiple regression analysis.

	Sum of Squares (SS)	**Degrees of freedom**	**Mean Square (MS)**
Model	$\mathbf{y}^T\mathrm{X}\widehat{\boldsymbol{\beta}} - n\bar{y}^2$	p	SS_{Model}/p
Error	$\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathrm{X}\widehat{\boldsymbol{\beta}}$	$n-p-1$	$SS_{Error}/(n-p-1)$
Total	$SS_{Model} + SS_{Error}$	$n-1$	

Testing hypotheses for:

1. **The overall model fit**

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$;

H_1: $\exists_j, j = 1, \ldots, p: \beta_j \neq 0$.

The test statistic is given by:

$$F = \frac{MS_{Model}}{MS_{Error}} \tag{2. 10}$$

Under the null hypothesis, F follows a F-Snedecor distribution with $p, n - p - 1$ degrees of freedom. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $F_0 \geq F_{(p,n-p-1,1-\alpha)}$, which represents the $1 - \alpha$ quantile of a F-Snedecor distribution with $p, n - p - 1$ degrees of freedom.

2. **The $\boldsymbol{\beta_j}$ coefficient**

H_0: $\beta_j = c$;

H_1: $\beta_j \neq c$.

The test statistic is given by:

$$T = \frac{\widehat{\beta_J} - c}{S\sqrt{d_{ii}}} \tag{2. 11}$$

where d_{ii} is (i, i)th element of the $(\mathrm{X^T X})^{-1}$ matrix. Under the null hypothesis, T follows a Student's t distribution with $n - p - 1$ degrees of freedom. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $|T_0| \geq t_{(n-p-1,1-\alpha/2)}$, which represents the $1 - \alpha/2$ quantile of a Student's t distribution with $n - p - 1$ degrees of freedom.

3. **A linear combination of $\boldsymbol{\beta}$s**

H_0: $\mathbf{a^T \beta} = \mathbf{c}$;

H_1: $\mathbf{a^T \beta} \neq \mathbf{c}$.

The test statistic is given by:

$$T = \frac{\mathbf{a^T \widehat{\beta} - c}}{S\sqrt{\mathbf{a^T (X^T X)^{-1} a}}} \tag{2. 12}$$

where $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta} = \mathrm{E}[\mathrm{Y}|X_1 = x_1, \dots, X_p = x_p]$. Under the null hypothesis, T follows a Student's t distribution with $n - p - 1$ degrees of freedom. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $|T_0| \geq t_{(n-p-1,1-\alpha/2)}$, which represents the $1 - \alpha/2$ quantile of a Student's t distribution with $n - p - 1$ degrees of freedom.

4. The comparison of nested models

Considering a complete model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p + \varepsilon$, it is possible to test if the reduced model can be $Y = \beta_0 + \beta_1 \beta X_1 + \cdots + \beta_k X_k + \varepsilon$, i.e.:

H_0: $\beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$;

H_1: $\exists_j, j = k+1, \dots, p: \beta_j \neq 0$.

The test statistic is given by:

$$F = \frac{(SS_{ErrorReducedModel} - SS_{ErrorFullModel})/(p-k)}{S_{FullModel}/(n-p-1)} \qquad (2.\ 13)$$

Under the null hypothesis, F follows a F-Snedecor distribution with $p - k, n - p - 1$ degrees of freedom. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $F_0 \geq F_{(p-k,n-p-1,1-\alpha)}$, which represents the $1 - \alpha$ quantile of a F-Snedecor distribution with $p - k, n - p - 1$ degrees of freedom.

Determination coefficient

The determination coefficient is a statistical calculation that represents the fraction of the y_i variability explained by the regression model, i.e., it measures how well its predictions approximate the real data points (Shieh, 2008).

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}} \qquad (2.\ 14)$$

If $R^2 = 1$, the regression predictions perfectly fit the data.

Due to the inflation that can be experienced by R^2 as more independent variables are added to the model, some authors recommend the use of an alternate but identically interpreted version of this measure, regularly referred to as adjusted R^2 ($R^2{}_{adj}$). It penalizes the statistic as extra variables are included in the model, and its value will always be less than or equal to that of R^2. It is computed as:

$$R^2{}_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \qquad (2.\ 15)$$

Model selection

A multiple regression model attributes a coefficient for each independent variable, meaning it contains all possible simpler models as special cases. There is usually an interest in selecting the most parsimonious model, i.e., a model that accomplishes a desired level of explanation or prediction with as

few explanatory variables as possible. One of the methods used to achieve this is the stepwise procedure, for which there are three main approaches:

- Forward selection: Starting with a model with no independent variables, the addition of each of them is tested using a selection criterion, only adding the variable whose inclusion gives the most statistically significant improvement of the fit. This is repeated until no variable is able to improve the model;
- Backward selection: Starting with a model containing all the candidate variables, the deletion of each of them is tested using a selection criterion, only deleting the variable whose exclusion gives the most statistically significant improvement of the fit. This is repeated until no further variables can be deleted without a statistically significant loss of fit;
- Bidirectional selection: A combination of the previous two, testing at each step whether variables should be included or excluded.

An automation of the stepwise selection is available through the *stepAIC()* function from the MASS package. This function is based on one of the most recognized selection criteria, the Akaike Information Criterion (AIC), an estimator of the relative quality of statistical models for a given set of data.

$$AIC = 2p - 2\ln(\hat{L}) \tag{2. 16}$$

where $\hat{L}$ is the value of the likelihood function for the model. The smaller the AIC, the better the model.

The direction argument of *stepAIC()* allows the specification of whether the process should only add terms ("forward"), only remove terms ("backward"), or do either as needed ("both").

Discordant observations

Besides checking if the selected model verifies the regression method's assumptions, it is also important to analyze the discordant observations that may be present in the data and possibly influencing the fitted model. There are two types of discordant observations:

1. **Outliers**, or atypical observations, are data points that appear to prominently differ from the rest. They can exist due to variability in the measurement or measurement errors, although a small number of outliers is to be expected in large samples, and not immediately attributed to abnormal circumstances.

   This assumption can be validated through the studentized (or jackknife) residuals (T_i).

   H_0: The ith observation is not an outlier;

   H_1: The ith observation is an outlier.

   The test statistic is given by:

$$T_i = r_i \left(\frac{n-p-1}{n-p-{r_i}^2}\right)^{1/2} \tag{2. 17}$$

where $r_i = \frac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}$ represents the ith observation's standardized residual (h_{ii} is the ith observation leverage). Under the hypothesis $\varepsilon \cap N(0, \sigma^2\mathbf{I})$, T_i follows an approximate Student's t distribution with $n-p-2$ degrees of freedom. For a significance level of α, the null hypothesis is rejected if the observed value of the test statistic $|T_0| \geq t_{(n-p-2,1-\alpha/2)}$, which represents the $1-\alpha/2$ quantile of a Student's t distribution with $n-p-2$ degrees of freedom.

The R function *outlierTest()*, from the car package, tests if the observation with the largest absolute value of studentized residuals is an outlier and calculates a corrected p-value.

2. **Influential observations** are data points whose deletion has a large effect on the parameter estimates, and therefore have the ability to change the fitting of the model (Everitt, 1998). One of the methods to identify influential observations is the Cook's distance:

$$D_i = \left(\frac{{r_i}^2}{p+1}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right) \quad (2.\ 18)$$

A criterion to consider an observation influential is for its Cook's distance to be bigger than 0.5.

2.8 Fisher's exact test

Fisher's exact test is a non-parametric statistical technique used to assess the hypothesis of independence of two categorical variables with two groups each, disposed in a 2×2 contingency table with fixed margins. This test is valid for all sample sizes, although it is usually applied as an alternative to the Chi-squared test of independence, when its size requirements are violated or when the data are very unequally distributed among the cells of the table. For two categorical variables X and Y:

H_0: There is no association between X and Y;

H_1: There is an association between X and Y.

Despite being originally formulated for 2×2 contingency tables, Fisher's test can be generalized to accommodate any $m * n$ contingency table, $m, n \geq 2$ (Mehta & Patel, 1983; Weisstein, n.d.). Assuming X and Y have m and n groups, respectively, the contingency table that represents their relationship has a $m * n$ dimension, where a_{ij} is the observation in the ith row and jth column ($i = 1, \dots, m$ and $j = 1, \dots, n$). The row and column sums are given by:

$$N = \sum R_i = \sum C_j \quad (2.\ 19)$$

The conditional probability of obtaining the present matrix given these row and columns sums is:

$$P_{cutoff} = \frac{(R_1!R_2!\dots R_m!)(C_1!C_2!\dots C_n!)}{N!\prod_{i,j} a_{ij}!} \quad (2.\ 20)$$

The significance level is then computed by summing the conditional probabilities of all the tables that have these same row and column sums that are no larger than P_{cutoff}.

The function used in R was *fisher.test()* from the stats package.

2.9 Multiple factor analysis of mixed data

Multivariate data analysis concerns all the statistical techniques that analyze multiple subjects' measurements simultaneously, i.e., data originated from two or more outcome variables. These variables can be either numerical, often handled by a Principal Component Analysis (PCA), or categorical, by a Multiple Correspondence Analysis (MCA).

Multiple factor analysis (Escofier & Pagès, 1994; Abdi et al., 2013) is a multivariate analysis method where variables are structured into conceptual groups (i.e., a set of variables measured for a sample of wines may contain groups of variables such as "odor", "taste" and "origin"). This method can be considered as an extension of PCA for categorical variables, MCA for numerical variables and Factor Analysis of Mixed Data (FAMD) where the active variables are of both types. Unlike regular Factor Analysis, the main idea of MFA is to give groups the same importance, by weighting each variable with the inverse of the variance of the first principal component of the group it belongs to. In standard MFA, the nature of the variables can vary from one group to another but not within groups. The multiple factor analysis of mixed data procedure allows this to take place.

Let n be the total number of observations in a dataset and p the total number of variables which describe them, separated in G groups. Each group is represented by a data matrix $\mathrm{X}^{(g)} = [{\mathrm{X}_1}^{(g)}\ {\mathrm{X}_2}^{(g)}]$, $g = 1, \dots, G$, where ${\mathrm{X}_1}^{(g)}$ contains the numerical variables of the gth group and ${\mathrm{X}_2}^{(g)}$ its categorical variables. The numerical columns of the matrices $\mathrm{X}^{(g)}$ are concatenated in a global numerical matrix $\mathrm{X}_1 = \left[{\mathrm{X}_1}^{(1)} \dots {\mathrm{X}_1}^{(G)}\right]$, of dimension $n \times b$ (b is the total number of numerical variables). The same is done for the categorical columns, in a global categorical matrix $\mathrm{X}_2 = [{\mathrm{X}_2}^{(1)} \dots {\mathrm{X}_2}^{(G)}]$, of dimension $n \times c$ (c is the total number of categorical variables). The total number of levels of the c categorical variables is m, and the total number of individuals belonging to level k ($k = 1, \dots, m$) is n_k.

Let $\mathrm{Z} = [\mathrm{Z}_1\ \mathrm{Z}_2]$, where Z_1 ($n \times b$) is the standardized version of X_1 (as in a regular PCA) and Z_2 ($n \times m$) is the centered version of the indicator matrix of X_2 (as in a regular MCA). Z has n rows and $b + m$ columns, where $b = b^{(1)} + \cdots + b^{(G)}$ and $m = m^{(1)} + \cdots + m^{(G)}$, given that $b^{(g)}$ and $m^{(g)}$ represent the number of numeric variables and of categorical variables' levels present in the gth group, respectively ($g = 1, \dots, G$). Let $\mathrm{N} = \frac{1}{n}\mathbb{I}_n$ be the diagonal matrix of order n of the weights of the rows of Z and $\mathrm{M} = diag(1, \dots, 1, \frac{n}{n_1}, \dots, \frac{n}{n_m})$ the diagonal matrix of order $b + m$ of the weights of the columns of Z.

Step 1 – **weighting the columns of Z to balance the importance of the groups:**

For $g = 1, \dots, G$, the first eigenvalue ${\lambda_1}^{(g)}$ of a PCA applied to $\mathrm{X}^{(g)}$ is computed. Then, the diagonal matrix P of the weights $\frac{1}{{\lambda_1}^{(t_k)}}$ can be obtained, where $t_k \in \{1, \dots, G\}$ denotes the group of the kth column of Z. Finally, the diagonal matrix MP representing the new weights of the columns of Z is calculated.

Step 2 – processing the factor coordinates (principal component scores):

The Generalized Singular Value Decomposition (GSVD) of Z with metrics $\mathrm{N} \in \mathbb{R}^n$ and $\mathrm{MP} \in \mathbb{R}^{b+m}$ gives

$$Z = U\Lambda V^T \qquad (2.\ 21)$$

where $\Lambda = diag(\sqrt{\lambda_1} \ldots \sqrt{\lambda_r})$ is the $r \times r$ diagonal matrix of the singular values of ZMZ^TN and Z^TNZM (r corresponds to the maximum number of linearly independent columns of Z), U is the $n \times r$ matrix of the first r eigenvectors of ZMZ^TN such that $U^TNU = \mathbb{I}_r$, and V is the $(b + m) \times r$ matrix of the first r eigenvectors of Z^TNZM such that $V^TMV = \mathbb{I}_r$ ($\mathbb{I}_r$ is the identity matrix of size r).

The $n \times r$ matrix of the factor coordinates of the n observations is given by $F = U\Lambda$. Therefore, the $(b + m) \times r$ matrix of the factor coordinates of the b quantitative variables and m levels is given by $A^* = MV\Lambda$, where the first b rows contain the factor coordinates of the numerical variables and the following m rows contain the factor coordinates of the categorical variables' levels.

Step 3 – squared loading processing:

The squared loadings symbolize the contribution of each of the p variables to the variance of the r principal components, i.e., the columns of F. Knowing $Var(f_i) = \|a_i\|_{MP}{}^2$, where f_i is the ith principal component and a_i is the ith loadings vector (columns of $A = \Lambda V$), the contribution of the jth variable $x_j, j = 1, \ldots, p$ to the variance of the ith principal component is:

$$\begin{cases} c_{ji} = \frac{1}{\lambda_1{}^{(t_j)}} a_{ji}{}^2 & \text{if } x_j \text{ is numerical;} \\ c_{ji} = \sum_{s \epsilon I_j} \frac{1}{\lambda_1{}^{(t_s)}} \frac{n}{n_s} a_{si}{}^2 & \text{if } x_j \text{ is categorical.} \end{cases} \qquad (2.\ 22)$$

where I_j represents the set of indices of the levels of the categorical variable x_j.

2.9.1 The PCAmixdata package

In order to overcome the previously mentioned limitations of standard PCA, MCA and MFA, the R package PCAmixdata (Chavent et al., 2017) comprises three main functions: *PCAmix()*, a PCA of a mix of numerical and categorical variables, *PCArot()*, a rotation following *PCAmix()*, and *MFAmix()*, a MFA of mixed multi-table data. These functions make no distinction between ordinal and nominal variables. While PCA for mixed data can be found in other R packages, this is the only existing implementation of a rotation for a PCA of mixed data and a MFA of mixed data. This package also proposes functions to plot graphical outputs, predict scores for new observations of the principal components of *PCAmix()*, *PCArot()* and *MFAmix()*, and project supplementary variables/groups of variables or levels on *PCAmix()*/ *MFAmix()* maps.

Chapter 3

Results

The dataset used in this project consists of the responses of 260 CEMG Lisbon County workers to a questionnaire applied between November 2012 and June 2013. Following a descriptive analysis, focused on the two versions of the CDRISC scale and some of the remaining collected variables, a preliminary selection of the best resilience scale score and the resilience predictor candidates is conducted, through the Spearman rank correlation coefficient and the Mann-Whitney and Kruskal-Wallis tests. After this, a regression analysis is applied, where resilience represents the dependent variable and the previously selected variables represent its predictors. Regression's assumptions are validated for the final model. Finally, in order to better explore the structure of the pre-selected variables while accounting for their types and the conceptual groups they may form, a multiple factor analysis of mixed data is conducted.

3.1 Exploratory analysis

The most important variables for the study's objectives are the resilience scale scores, which were calculated with the full 25 scale items (CDRISC-25) and also with 10 of the 25 items (CDRISC-10).

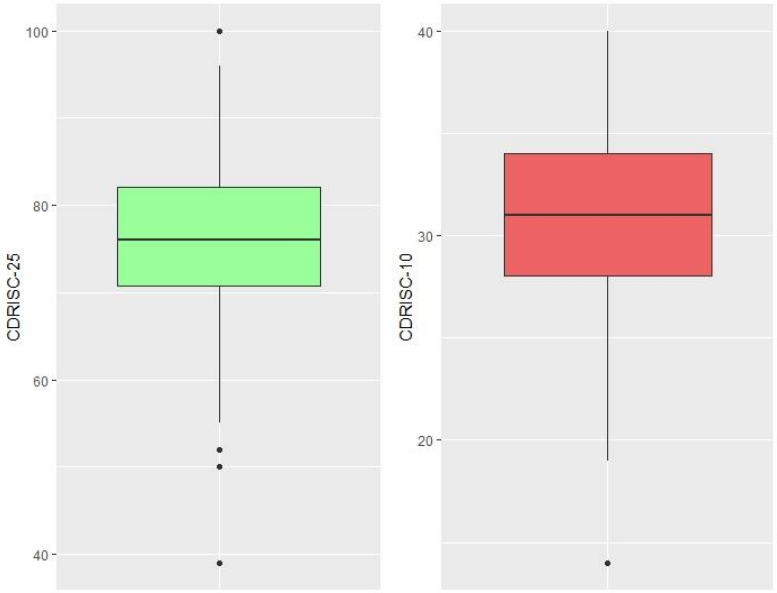

Figure 3.1.1: Boxplots of CDRISC-25 (left) and CDRISC-10 (right).

As Figure 3.1.1 shows, both boxplots' inter-quartile ranges seem to be almost symmetrical. The boxplot for CDRISC-10 appears to have a longer bottom whisker, which means the lower 10-item resilience scores have a lot of variability. It is also worthy to note that CDRISC-25's boxplot contains four outlier candidates, while CDRISC-10 only presents one. This might be an indication that CDRISC-10 is a more appropriate resilience scale score for the analyzed workers than CDRISC-25.

Table 3.1.1: Summary of some of the questionnaire's numerical variables.

Variable	n	Minimum	Median	Mean	Maximum	Standard deviation
Absenteeism	256	-6	6.50	7.59	42	6.29
Number of missed days in the last 12 months	258	0	0	4.55	150	14.58
Age	260	22	42	42.26	64	8.46
Impact of the economic crisis in a scale of 0 to 10	260	0	6	5.25	10	3.03
Job Satisfaction Scale	260	1	4	4.43	7	1.08
Presenteeism	260	20	80	76.46	100	13.43
Subjective Happiness Scale	260	2	5	4.70	7	0.78
CDRISC-25	260	39	76	75.88	100	8.75
CDRISC-10	260	14	31	30.68	40	4.10
Neutrophils	260	34	56	57.02	82	8.11
Lymphocytes	260	11	32	32.28	58	7.57
Monocytes	260	2	7	7.43	15	1.68
Eosinophils	260	0	2	2.88	11	1.90
Basophils	260	0	1	0.67	3	0.52
Platelets	260	102	228	231.71	461	50.46
Glucose	260	50	93	95.47	378	22.76
Urea	260	14	31	32.07	53	7.44
Creatinine	260	0	1	0.99	1	0.11
Total Cholesterol	260	119	200.50	204.48	316	35.89
High-density lipoprotein (HDL)	260	25	57	58.10	108	16.41
Low-density lipoprotein (LDL)	260	52	127	132.35	231	33.81
Triglycerides	260	28	85	102.61	461	59.30
BMI	260	18.20	25.60	25.96	45.80	4.15
Systolic blood pressure	260	81	115.50	116.50	167	14.55
Diastolic blood pressure	260	54	77.50	78.22	114	9.99
Heart rate	260	46	69	69.88	104	10.50
Cardiovascular risk	260	0	3	3.14	7	1.50

From Table 3.1.1, the average respondent is about 42 years old, works almost 8 hours longer than the amount set in their contract, has missed work for approximately 4 days in the previous year and presents 24.5% of perceived performance loss. In average, the subjects in this sample are pre-obese and have a total cholesterol level higher than the reference level.

Table 3.1.2: Summary of some of the questionnaire's categorical variables.

Variable	Frequency	%
Psychological distress (MHI-5)		
Yes	53	20.4%
No	207	79.6%
Medication for chronic anxiety in the previous 2 weeks		
Yes	33	12.7%
No	227	87.3%
Interests/hobbies		
Yes	237	91.5%
No	21	8.1%
Doesn't know	1	0.4%
Work relationships (ASSET)		
Very low levels of the stressor	29	11.2%
Low levels of the stressor	40	15.4%
Average levels of the stressor	176	67.7%
High levels of the stressor	7	2.7%
Very high levels of the stressor	8	3.1%
Overload (ASSET)		
Very low levels of the stressor	21	8.1%
Low levels of the stressor	21	8.1%
Average levels of the stressor	183	70.4%
High levels of the stressor	19	7.3%
Very high levels of the stressor	16	6.2%
Job security (ASSET)		
Very low levels of the stressor	17	6.5%
Low levels of the stressor	10	3.8%
Average levels of the stressor	211	81.2%
High levels of the stressor	15	5.8%
Very high levels of the stressor	7	2.7%
Control (ASSET)		
Very low levels of the stressor	34	13.1%
Low levels of the stressor	50	19.2%
Average levels of the stressor	149	57.3%
High levels of the stressor	18	6.9%
Very high levels of the stressor	9	3.5%
Resources and communication (ASSET)		
Very low levels of the stressor	21	8.1%
Low levels of the stressor	55	21.2%
Average levels of the stressor	169	65.0%
High levels of the stressor	13	5.0%
Very high levels of the stressor	2	0.8%
Aspects of the job (ASSET)		
Very low levels of the stressor	11	4.2%
Low levels of the stressor	46	17.7%
Average levels of the stressor	189	72.7%

High levels of the stressor	9	3.5%
Very high levels of the stressor	5	1.9%
Perceived commitment of employee to organization (ASSET)		
Very low levels of commitment	1	0.4%
Low levels of commitment	2	0.8%
Average levels of commitment	98	37.7%
High levels of commitment	51	19.6%
Very high levels of commitment	108	41.5%
Physical health (ASSET)		
Very good health levels	34	13.1%
Good health levels	41	15.8%
Average health levels	165	63.5%
Low health levels	12	4.6%
Very low health levels	8	3.1%
Psychological wellbeing (ASSET)		
Very good health levels	25	9.6%
Good health levels	45	17.3%
Average health levels	163	62.7%
Low health levels	17	6.5%
Very low health levels	10	3.8%
Productivity		
100%	36	13.8%
90-99%	105	40.4%
80-89%	79	30.4%
70-79%	26	10.0%
<70%	14	5.4%
Depression		
Yes	56	21.7%
No	179	69.4%
Doesn't know	23	8.9%
Chronic anxiety		
Yes	21	8.1%
No	213	82.6%
Doesn't know	24	9.3%
Medication for chronic anxiety		
Yes	19	7.4%
No	88	34.2%
Doesn't apply	146	56.8%
Doesn't know	4	1.6%

From Table 3.1.2, over 90% of the workers have interests/hobbies and over 50% have perceived productivity between 90% and 100%. The majority of the subjects present average levels of the stressors, very high levels of perceived commitment of employee to organization and average physical and psychological health levels. As far as medical conditions go, more than half of the respondents do not suffer from depression or chronic anxiety.

3.2 Choosing the dependent variable

The original CDRISC scale score, obtained from the entirety of the scale's 25 items (CDRISC-25), and the abbreviated version, only containing 10 out of the 25 items (CDRISC-10), were both computed for our sample. The first thing to assess was which of these two variables should be used as the dependent variable throughout the analysis.

In order to test if there was a significant association between both scale scores, and due to the fact the bivariate normality assumption related to the Pearson correlation coefficient was not verified, the Spearman rank correlation coefficient was instead calculated. Its observed value of 0.85 denotes a high positive association between the variables ($p < 0.00001$). This is also made evident by the scatter plot in Figure 3.2.1.

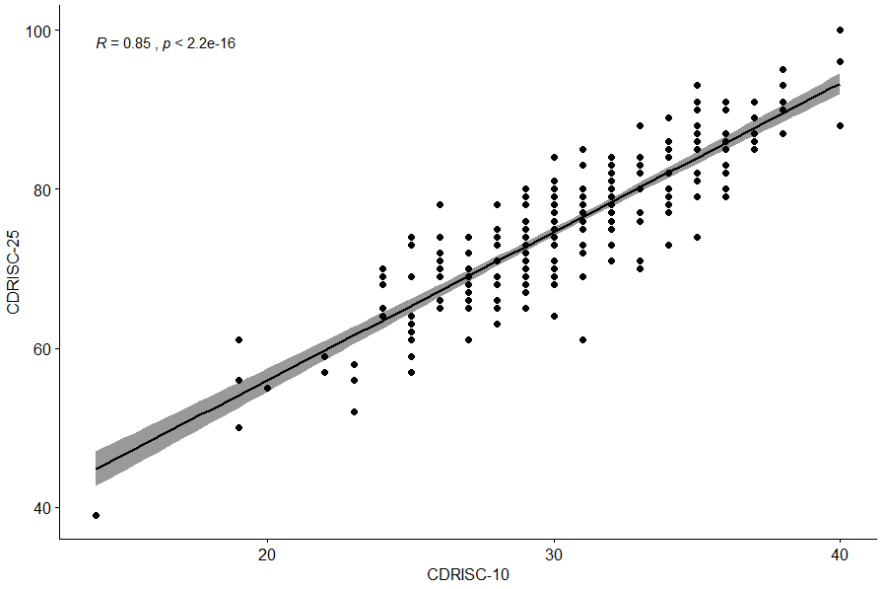

Figure 3.2.1: Scatter plot of CDRISC-10 *vs* CDRISC-25.

Given the high association between CDRISC-25 and CDRISC-10, it is clear to presume that the original 25 question scale does not present an advantage over its abbreviated counterpart, which makes it redundant. For these reasons, CDRISC-10 was chosen as the dependent variable in this study.

3.3 Choosing the independent variables

Before applying a regression analysis, there was a need to reduce the large number of predictor candidates (103, i.e., the 106 variables in the dataset minus the two resilience scale scores and the subjects' ID). Therefore, a manual pre-selection of variables was firstly conducted.

3.3.1 Numerical variables

To select the numerical variables (41 in total, detailed in Appendix A) that may be relevant to explain resilience, the Spearman correlation coefficient was calculated between CDRISC-10 and each variable of this type present in the dataset. After correcting p-values using the Benjamini-Hochberg method, the variables whose observed r_s is significantly different from 0 are:

Table 3.3.1.1: Spearman's correlation coefficient estimates and the corrected p-values for the respective significant tests.

	Spearman's correlation coefficient (estimate)	**Corrected p-value**
Number of missed days	-0.18	0.03116
Subjective Happiness Scale	0.20	0.01761
Job Satisfaction Scale	0.24	0.00259
Presenteeism	0.28	0.00021

All the estimates present in Table 3.3.1.1 represent weak rank correlations, and therefore, weak monotonic relationships. The Spearman's coefficient estimate between resilience and the number of missed days is negative, which means resilience tends to decrease when the number of missed days increases. The remaining coefficient estimates are positive, which means resilience tends to increase when subjective happiness, job satisfaction and proportion of performance gain increase.

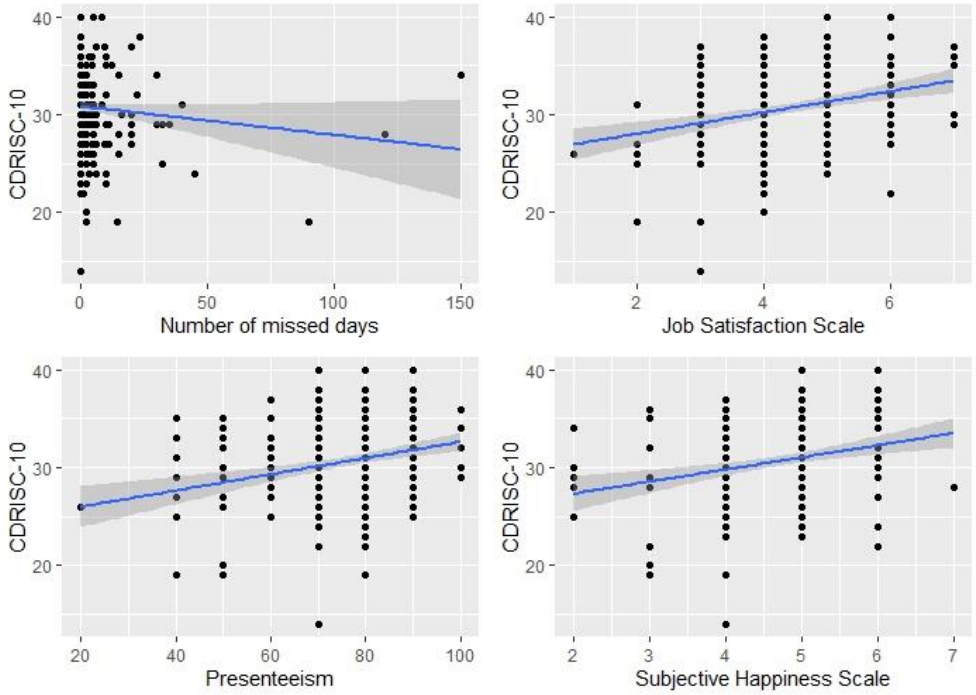

Figure 3.3.1.1: Scatter plots of CDRISC-10 *vs* the variables with significant Spearman's rank correlation coefficient.

By observing the plot for "Number of missed days" (Figure 3.3.1.1), it is clear that the fitted line is only supported by a few observations with high values of missed days, which explains the low $|r_s|$ and its high p-value. This suggests an absence of causality between this variable and resilience. The remaining plots present a similar behavior among themselves: very slight increase of resilience as the variables' values also increase.

Even though the Spearman correlation coefficients of these variables suggest weak associations, they were still chosen to integrate the regression analysis due to their significant p-values.

3.3.2 Categorical variables

To select the categorical variables (62 in total, detailed in Appendix A) that may be relevant to explain resilience, the Mann-Whitney and Kruskal-Wallis non-parametric tests were computed: the first between CDRISC-10 and categorical variables with exactly two groups and the latter between CDRISC-10 and categorical variables with more than two groups.

The variables that present statistically significant differences of resilience levels between groups are:

Table 3.3.2.1: Mann-Whitney and Kruskal-Wallis tests' corrected p-values.

Variable	Test	Corrected p-value
Psychological distress	Mann-Whitney	0.00001
Anxiety medication in the previous 2 weeks		0.02960
Psychological wellbeing (ASSET)	Kruskal-Wallis	< 0.00001
Productivity		0.00034
Chronic Anxiety		0.00047
Control (ASSET)		0.00109
Aspects of the Job (ASSET)		0.00205
Physical health (ASSET)		0.00375
Medication for chronic anxiety		0.00375
Interests/hobbies		0.00850
Work relationships (ASSET)		0.00850
Overload (ASSET)		0.00850
Perceived commitment of employee to organization (ASSET)		0.00850
Job Security (ASSET)		0.01769
Depression		0.02819
Resources and Communication (ASSET)		0.04584

The variables with significant p-values obtained from the application of Mann-Whitney and Kruskal-Wallis tests are mainly ASSET's subscales and variables related to psychological diseases. "Resources and Communication" presents the highest corrected p-value and "Psychological wellbeing" the lowest.

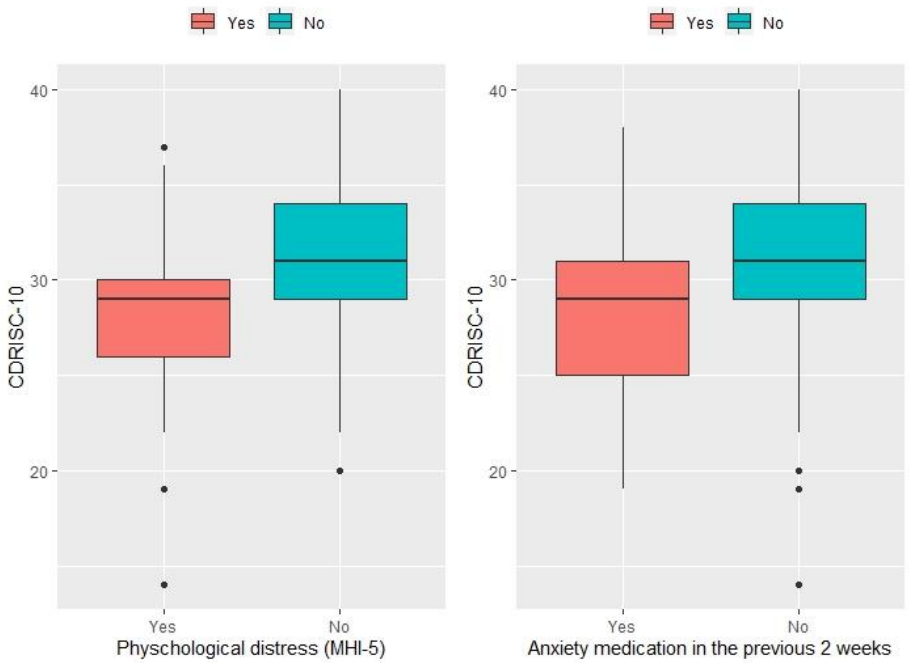

Figure 3.3.2.1: Parallel boxplots of CDRISC-10 *vs* the variables with significant Mann-Whitney's test.

For the variables with only two groups, "Psychological distress" and "Anxiety medication in the previous 2 weeks", the differences in resilience are observed between those groups. According to the graphs in Figure 3.3.2.1, the subjects that present psychological distress and who have taken anxiety medication in the previous 2 weeks have significantly less resilience than their counterparts.

However, for the variables with three or more groups, it is necessary to identify exactly which groups differ. These results are presented in the next section.

3.3.2.1 Multiple pairwise comparisons

The categorical variables with a significant p-value obtained from the application of the Kruskal-Wallis test were subjected to the Conover's post-hoc test, to discover which groups within each variable have significant differences in resilience levels. The p-value correction method was once again Benjamini-Hochberg.

Despite the fact some of these variables contain groups with very few elements, these groups were not eliminated or combined with other groups in order to maintain original structure of the data, specifically regarding the ASSET subscales, which are meant to have all five classes.

For the previously selected variables, the following groups differ significantly:

Interests/hobbies

- "Yes" and "No"

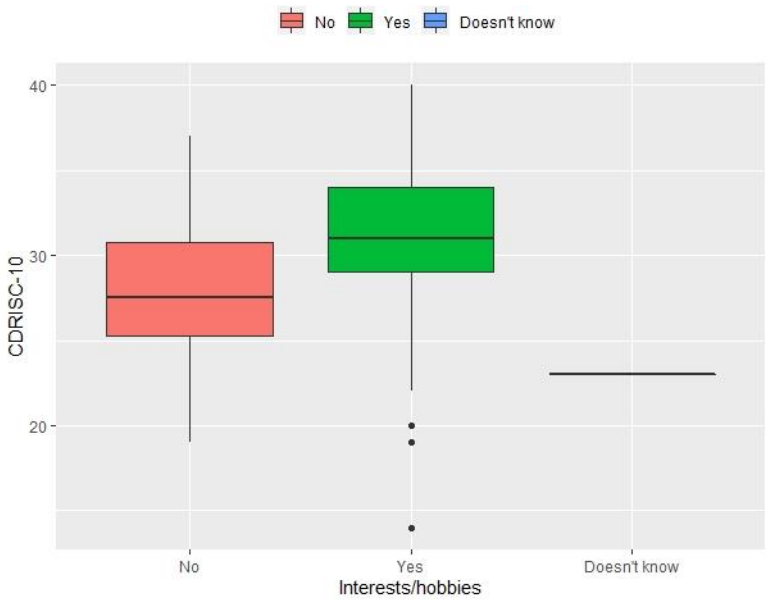

Figure 3.3.2.1.1: Parallel boxplots of CDRISC-10 *vs* Interests/hobbies.

According to Figure 3.3.2.1.1, subjects with no interests or hobbies appear to have lower levels of resilience than those who have them.

Work relationships (ASSET)

- "Very low" levels of the stressor and "Very high"/"Average" levels of the stressor
- "Low" levels of the stressor and "Very high"/ "Average" levels of the stressor

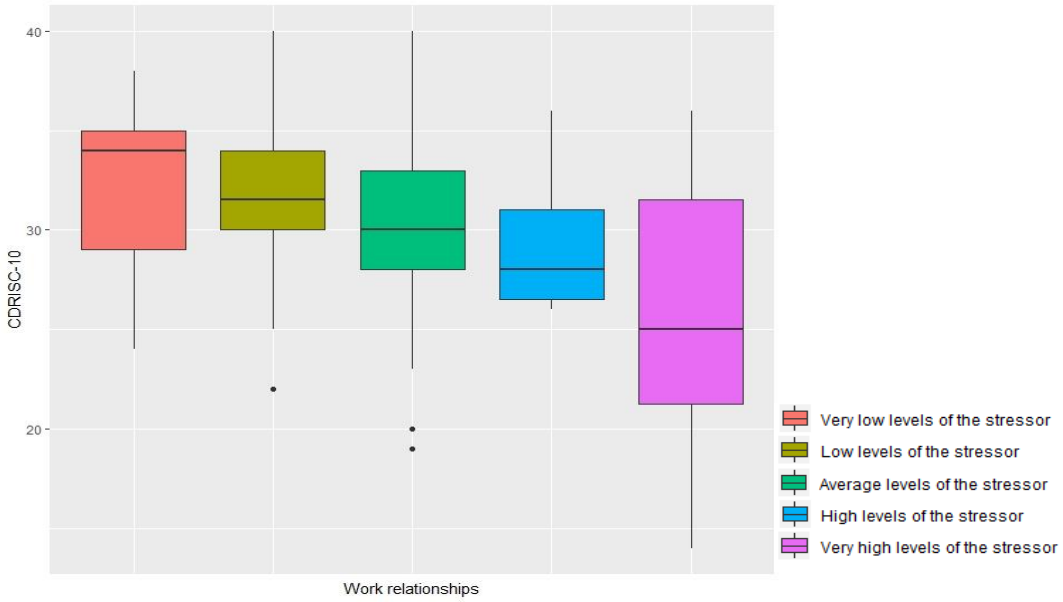

Figure 3.3.2.1.2: Parallel boxplots of CDRISC-10 *vs* Work relationships.

Figure 3.3.2.1.2 shows that the higher the levels of the stress caused by work relationships, the lower resilience tends to be. The significant differences related to “Average levels of the stressor” may exist due to the presence of outlier candidates.

Overload (ASSET)

- “Very low” levels of the stressor and “Very high”/ “High”/“Average” levels of the stressor
- “Low” levels of the stressor and “Very high”/ “High” levels of the stressor

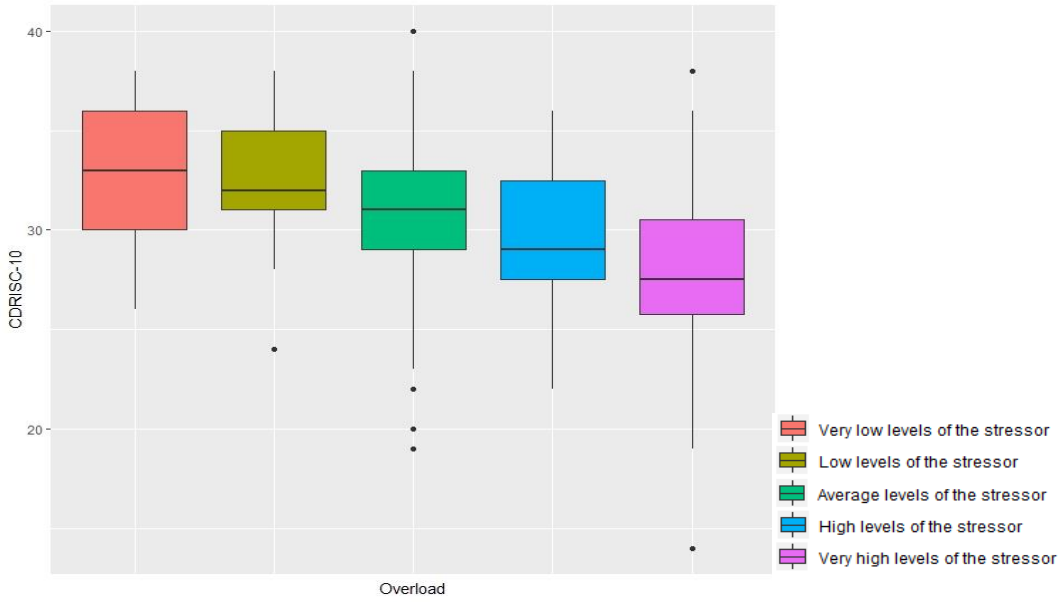

Figure 3.3.2.1.3: Parallel boxplots of CDRISC-10 *vs* Overload.

The parallel plots in Figure 3.3.2.1.3 seem to indicate that the higher the levels of stress caused by work overload, the lower resilience tends to be. The significant differences obtained by the post-hoc test are in line with the graphic representation.

Job security (ASSET)

- “Very low” levels of the stressor and “Very high”/”Average”/”Low” levels of the stressor
- “Very high” levels of the stressor and “Average” levels of the stressor

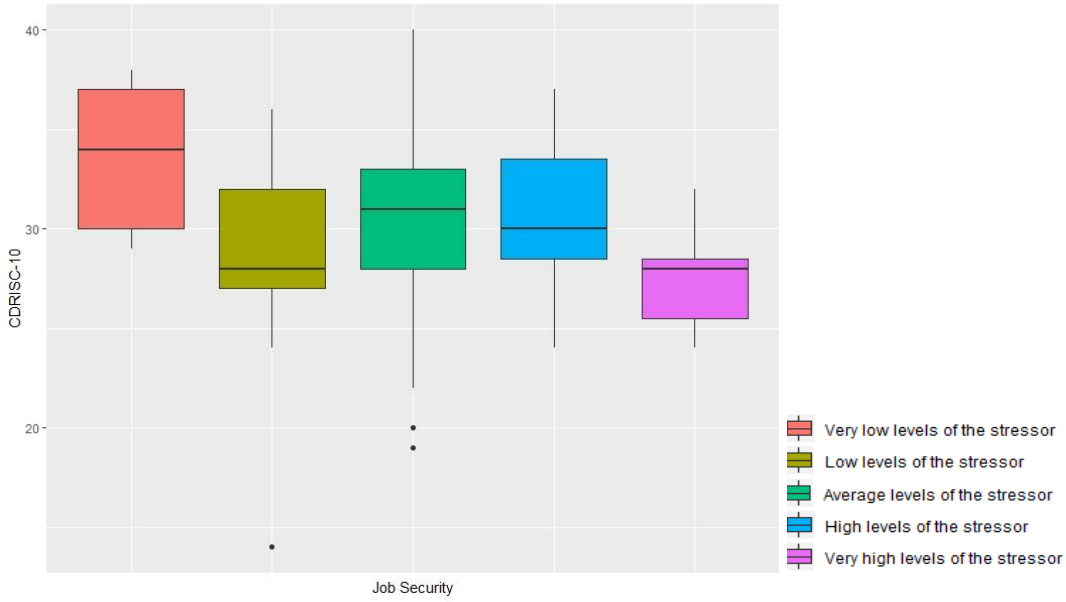

Figure 3.3.2.1.4: Parallel boxplots of CDRISC-10 *vs* Job security.

Regarding job security, subjects with very low levels of this stressor appear to have higher resilience than the rest. However, the fact that those with low levels of the stressor present lower resilience than the ones with average and high levels of the stressor is not to be expected. This is most likely due to the uneven group sizes and the outlier candidates.

Control (ASSET)

- "Very low" levels of the stressor and "High"/"Average"/"Low" levels of the stressor

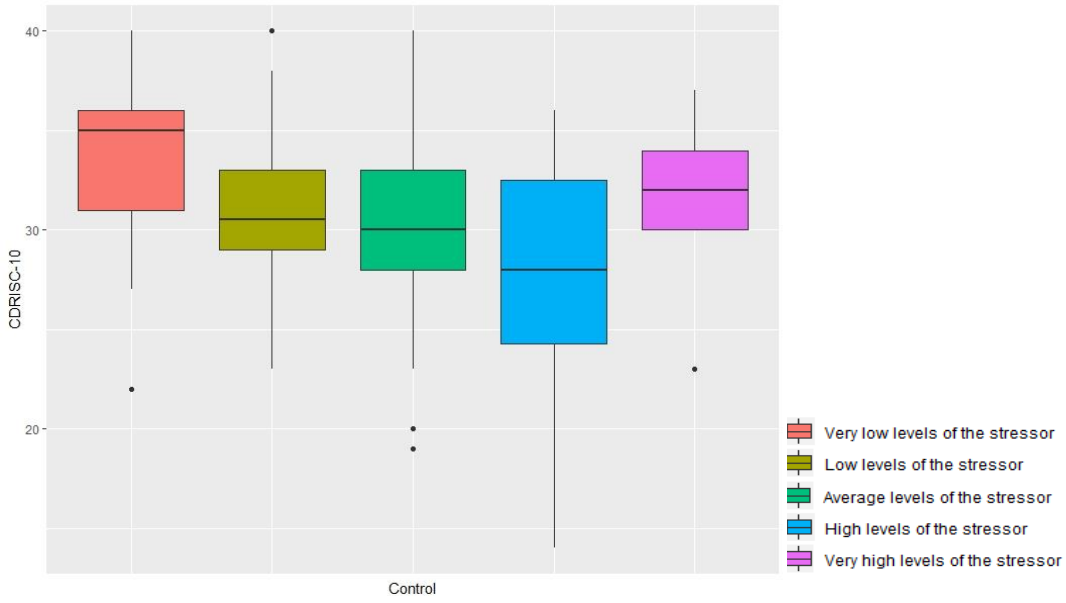

Figure 3.3.2.1.5: Parallel boxplots of CDRISC-10 *vs* Control.

By observing the boxplots in Figure 3.3.2.1.5, it is apparent that subjects with very low levels of the control stressor present higher amounts of resilience than the remaining. The fact that those with very high levels of this stressor present higher resilience than the ones with low, average and high levels of the stressor is not to be expected. This may be due to the same reasons presented in the comment of Figure 3.3.2.1.4.

Resources and Communication (ASSET)

- "Very low" levels of the stressor and "Very high"/"Average" levels of the stressor

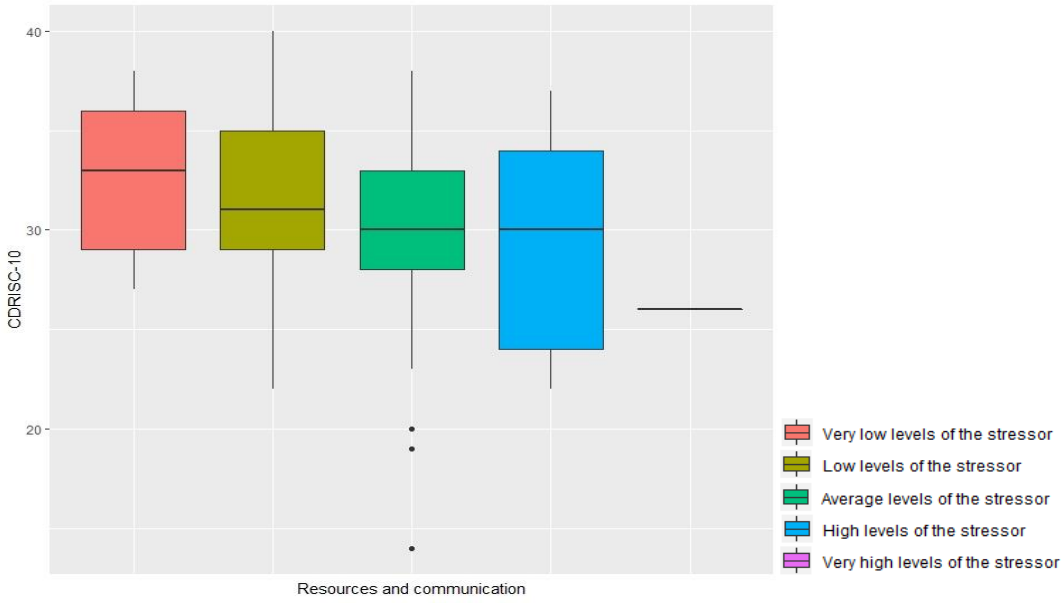

Figure 3.3.2.1.6: Parallel boxplots of CDRISC-10 *vs* Resources and communication.

Regarding the resources and communication stressor, the boxplots of Figure 3.3.2.1.6 show that the amount of resilience tends to decrease as the levels of this stressor increase. There are several outlier candidates in the group of subjects who have average levels of the stressor, and very few values in the group of subjects who have very high levels of the stressor, which may explain the results of the post-hoc test.

Aspects of the job (ASSET)

- "Very low" levels of the stressor and "Very high"/"High"/"Average" levels of the stressor
- "Low" levels of the stressor and "Very high"/"High"/"Average" levels of the stressor
- "Average" and "High" levels of the stressor

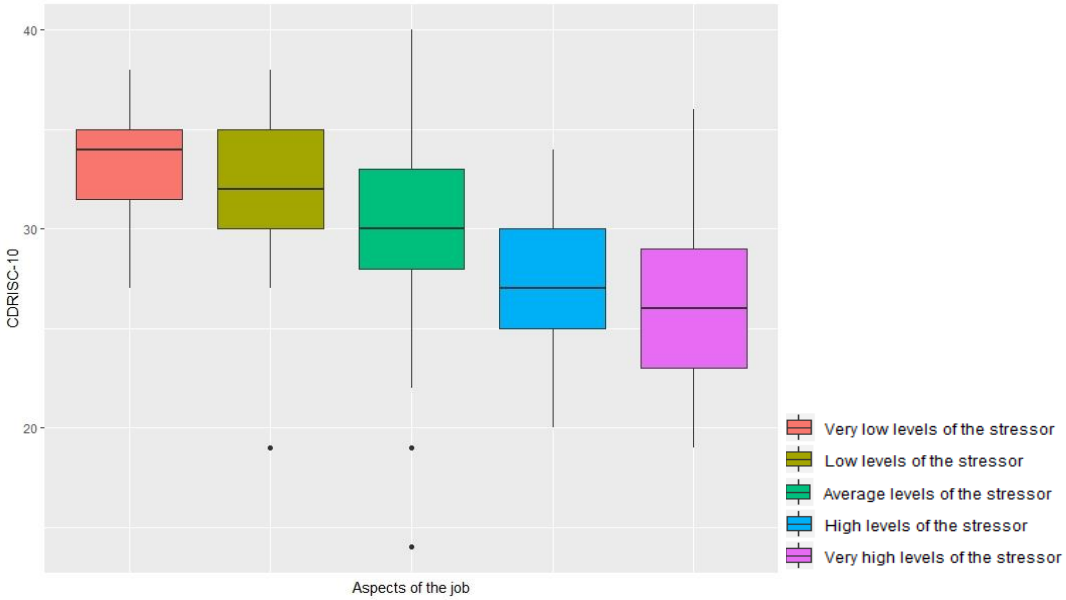

Figure 3.3.2.1.7: Parallel boxplots of CDRISC-10 *vs* Aspects of the job.

The boxplots of Figure 3.3.2.1.7 represent the expected behavior for a stressor: resilience increases as the levels of the stressor decrease. The results obtained by the post-hoc test seem to correspond in general to the graphic representation.

Perceived commitment of employee to organization (ASSET)

- "Very high" levels of commitment and "Average" levels of commitment

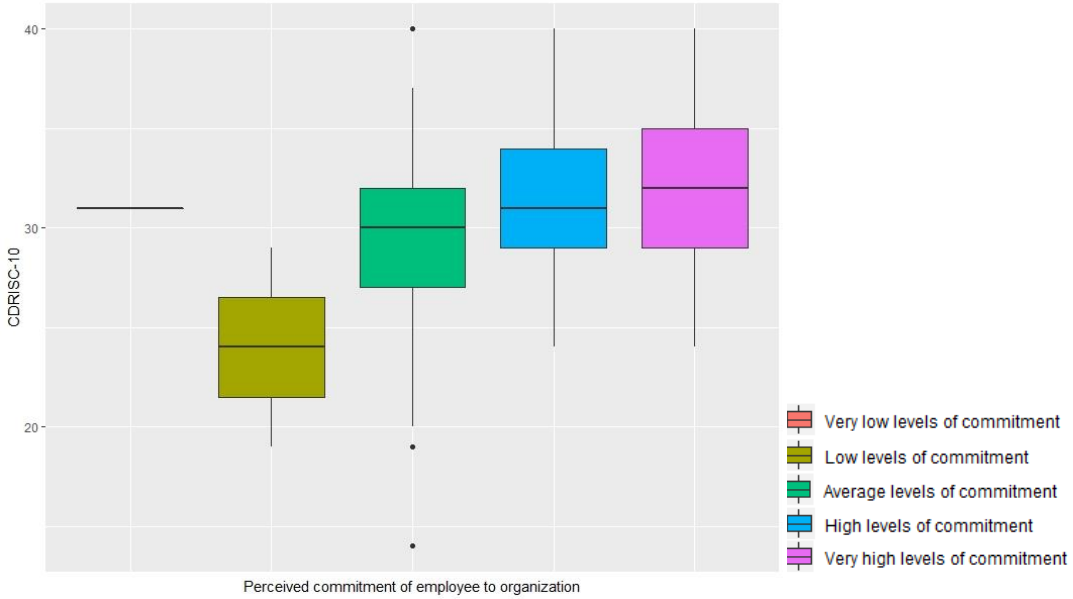

Figure 3.3.2.1.8: Parallel boxplots of CDRISC-10 *vs* Perceived commitment of employee to organization.

In general, subjects with higher levels of perceived commitment of employee to organization also have higher levels of resilience. This is not observed for the subjects with very low levels of commitment, who present higher resilience than those with low and average levels of commitment. This discrepancy is most likely also due to the uneven group sizes.

Physical health (ASSET)

- "Very good" health levels and "Average"/"Low"/"Very low" health levels
- "Very low" health levels and "Good"/"Average" health levels

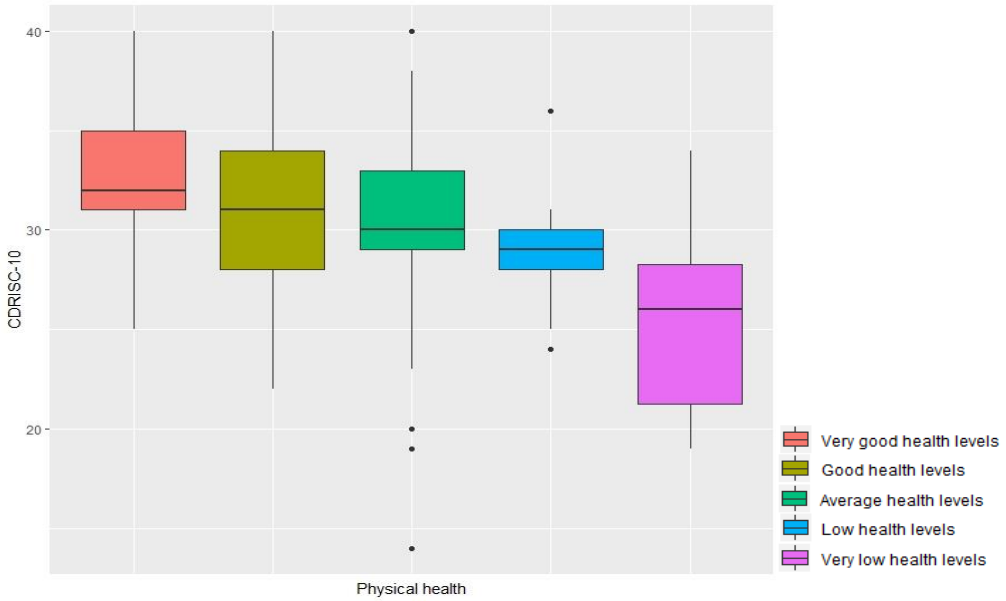

Figure 3.3.2.1.9: Parallel boxplots of CDRISC-10 *vs* Physical health.

Regarding physical health, resilience seems to increase as the level of health increases, which is to be expected.

Psychological wellbeing (ASSET)

- "Very good" health levels and "Good"/"Average"/"Low"/"Very low" health levels
- "Good" health levels and "Average"/"Low"/"Very low" health levels
- "Average" health levels and "Low"/"Very low" health levels

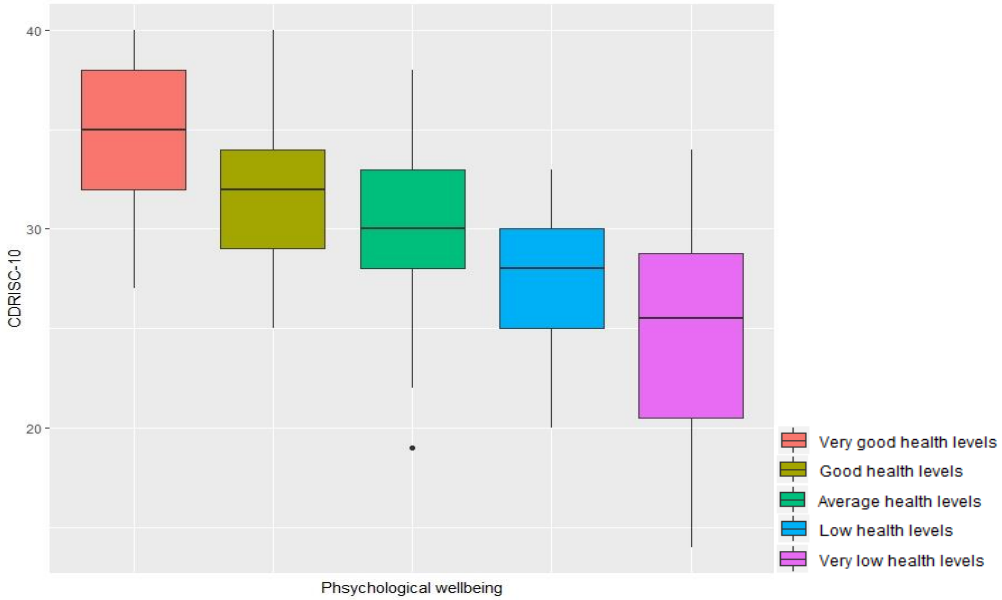

Figure 3.3.2.1.10: Parallel boxplots of CDRISC-10 *vs* Psychological wellbeing.

As with physical health, resilience seems to increase as psychological health levels increase. The results obtained by the post-hoc test are in line with the graphic representation.

Productivity

- "100%" and "90-99%"/"80-89%"/"70-79%"/"<70%"
- "90-99%" and "80-89%"/"70-79%"

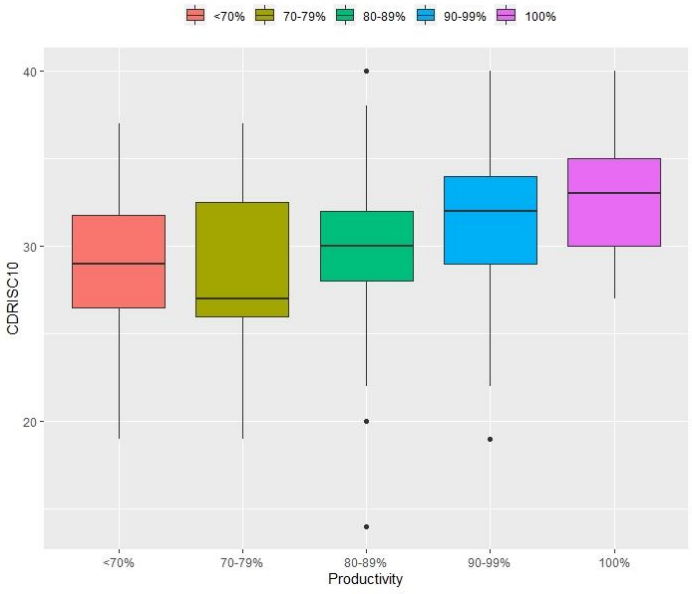

Figure 3.3.2.1.11: Parallel boxplots of CDRISC-10 *vs* Productivity.

Regarding productivity, the boxplots above suggest that, in general, subjects who have higher perceived work productivity also present higher resilience. A significant difference between 90-99% and <70% of productivity is to be expected but was not confirmed by the post-hoc test, which may be due to the outlier candidate of "90-99%".

Depression

- "Yes" and "No"

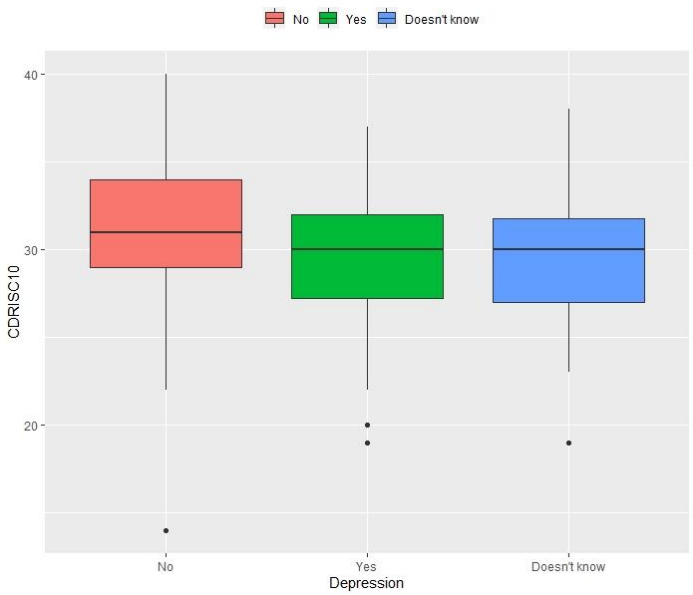

Figure 3.3.2.1.12: Parallel boxplots of CDRISC-10 *vs* Depression.

Subjects who have not had depression present slightly higher resilience than those who have or who do not know. It is possible that the significant difference obtained by the post-hoc test is due to the outlier candidates in the group of subjects with depression.

Chronic Anxiety

- "No" and "Yes"/"Doesn't know"

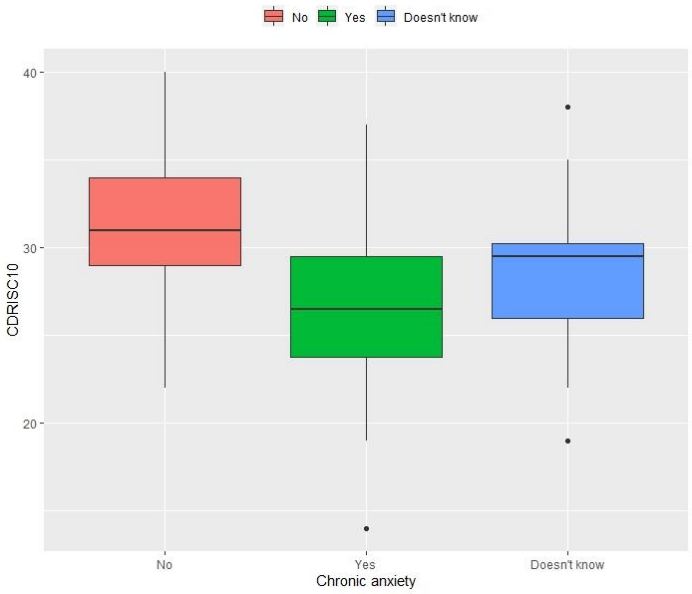

Figure 3.3.2.1.13: Parallel boxplots of CDRISC-10 *vs* Chronic anxiety.

According to the boxplots of Figure 3.3.2.1.13, subjects who have had chronic anxiety display lower resilience than those who have not had the disease or who do not know. The results obtained by the post-hoc test are in line with the graphic representation.

Medication for chronic anxiety

- "Yes" and "No"/"Doesn't apply"

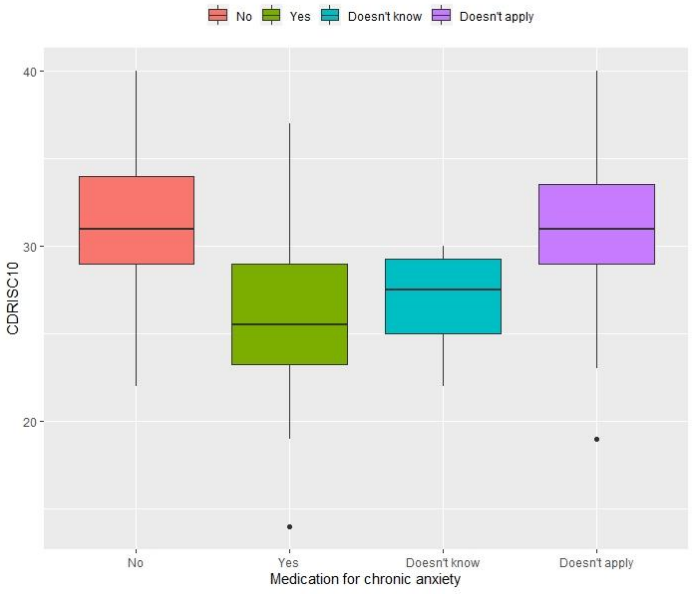

Figure 3.3.2.1.14: Parallel boxplots of CDRISC-10 *vs* Medication for chronic anxiety.

Workers who have not had chronic anxiety or who do not take medication for this condition show higher resilience than those who take medication and those who do not know. The graphic representation validates the results obtained by the post-hoc test.

3.4 Resilience prediction

A multiple linear regression was conducted to predict resilience (measured by CDRISC-10, the dependent variable), based on all the selected variables obtained from the significant Spearman rank correlation coefficients and the Mann-Whitney and Kruskal-Wallis tests (the independent variables). This first model, from now on referred to as **Model 1**, was found significant by the F-statistic ($F_{obs} = 3.35$, $p < 0.00001$), with an observed ${R^2}_{adj}$ of 0.34. This means the overall model fit is statistically significant and that this model explains 34% of the resilience's variance.

Table 3.4.1: Parameter estimates and their standard errors, observed values of the t-statistic and p-values for Model 1.

Model 1	Estimate	Standard error	Observed t-statistic	p-value
Intercept	31.99487	4.85546	6.58900	< 0.00001
Psychological distress (ref: Yes)				
No	0.67815	0.77419	0.87600	0.38212
Anxiety medication in the previous 2 weeks (ref: No)				
Yes	0.62073	0.93979	0.66000	0.50971
Interests/hobbies (ref: No)				
Yes	2.14467	0.88526	2.42300	0.01631
Doesn't know	-3.74897	3.67512	-1.02000	0.30893
Work relationships (ASSET) (Ref: Very low)				
Low levels of the stressor	0.55152	1.07950	0.51100	0.60999
Average levels of the stressor	0.16770	1.12672	0.14900	0.88183
High levels of the stressor	2.61038	2.18107	1.19700	0.23280
Very high levels of the stressor	-2.84124	2.03471	-1.39600	0.16416
Overload (ASSET) (Ref: Very low)				
Low levels of the stressor	-1.16791	1.21960	-0.95800	0.33942
Average levels of the stressor	-2.45706	1.16840	-2.10300	0.03674
High levels of the stressor	-3.41977	1.57360	-2.17300	0.03095
Very high levels of the stressor	-3.69755	1.51928	-2.43400	0.01583
Job security (ASSET) (Ref: Very low)				
Low levels of the stressor	-3.39831	1.47049	-2.31100	0.02186
Average levels of the stressor	0.66400	1.06586	0.62300	0.53402
High levels of the stressor	1.61372	1.42565	1.13200	0.25904
Very high levels of the stressor	-2.60165	1.99078	-1.30700	0.19278
Control (ASSET) (Ref: Very low)				
Low levels of the stressor	-1.64585	0.97522	-1.68800	0.09305
Average levels of the stressor	-1.16848	1.00048	-1.16800	0.24424
High levels of the stressor	-1.40687	1.47416	-0.95400	0.34107
Very high levels of the stressor	1.23838	1.76359	0.70200	0.48338

Resources and communication (ASSET) (Ref: Very low)				
Low levels of the stressor	0.58351	1.13191	0.51600	0.60677
Average levels of the stressor	0.83511	1.18677	0.70400	0.48245
High levels of the stressor	1.06280	1.75623	0.60500	0.54577
Very high levels of the stressor	-3.03409	3.41424	-0.88900	0.37527
Aspects of the job (ASSET) (Ref: Very low)				
Low levels of the stressor	0.06782	1.30089	0.05200	0.95847
Average levels of the stressor	-0.24418	1.32807	-0.18400	0.85431
High levels of the stressor	-2.35745	1.86584	-1.26300	0.20790
Very high levels of the stressor	-0.25480	2.47457	-0.10300	0.91809
Perceived commitment of employee to organization (ASSET) (Ref: Very low)				
Low levels of commitment	-1.54642	4.71744	-0.32800	0.74340
Average levels of commitment	-0.65382	3.62087	-0.18100	0.85689
High levels of commitment	-0.25639	3.65576	-0.07000	0.94416
Very high levels of commitment	0.62282	3.62919	0.17200	0.86392
Physical health (ASSET) (Ref: Very good)				
Good health levels	-0.29253	0.90130	-0.32500	0.74585
Average health levels	-0.09085	0.79291	-0.11500	0.90889
Low health levels	0.52795	1.47403	0.35800	0.72060
Very low health levels	-1.27861	1.90805	-0.67000	0.50357
Psychological wellbeing (ASSET) (Ref: Very good)				
Good health levels	-1.90219	0.95012	-2.00200	0.04664
Average health levels	-3.06914	0.92795	-3.30700	0.00112
Low health levels	-4.77014	1.59089	-2.99800	0.00306
Very low health levels	-4.18030	2.00489	-2.08500	0.03835
Productivity (Ref: 100%)				
90-99%	0.01522	0.74994	0.02000	0.98383
80-89%	-1.00802	0.86073	-1.17100	0.24296
70-79%	-0.60892	1.09242	-0.55700	0.57788
<70%	0.37590	1.46724	0.25600	0.79806
Depression (ref: No)				
Yes	-0.41223	0.67258	-0.61300	0.54064
Doesn't know	0.31772	1.07445	0.29600	0.76776
Chronic anxiety (ref: No)				
Yes	-0.85136	1.57420	-0.54100	0.58924
Doesn't know	-0.33419	1.24660	-0.26800	0.78892
Medication for chronic anxiety (ref: No)				
Yes	-2.32324	1.71557	-1.35400	0.17721
Doesn't apply	-0.88394	0.55870	-1.58200	0.11521
Doesn't know	0.04920	2.61810	0.01900	0.98502
Number of missed days	-0.01786	0.01807	-0.98900	0.32404

Job Satisfaction Scale	-0.38029	0.32572	-1.16800	0.24440
Presenteeism	0.04031	0.02175	1.85300	0.06537
Subjective Happiness Scale	0.24782	0.34805	0.71200	0.47729

According to the p-values in Table 3.4.1, having interests and hobbies differs significantly from not having them. Also, average, high and very high levels of the overload stressor differ significantly from very low levels of this stressor. This can be observed as well between low and very low levels of the job security stressor. Finally, good, average, low and very low psychological wellbeing levels all differ significantly from very good health levels.

Due to these results and the fact that the number of variables in **Model 1** is still quite large, the stepwise model selection method was applied in order to find the most parsimonious model.

Table 3.4.2: Stepwise selection's final models.

Method	Final model (simplified)	Observed R^2_{adj}	AIC
	Model 1.1		
Bidirectional (starting with Model 1)	CDRISC-10 = Interests/hobbies + Psychological wellbeing + Job security + Medication for chronic anxiety + Presenteeism + Overload + Work relationships + Control	0.3659	588.26
	Model 1.2		
Bidirectional (starting with the null model)	CDRISC-10 = Interests/hobbies + Psychological wellbeing + Job security + Medication for chronic anxiety + Presenteeism + Overload	0.3473	587.78

As Table 3.4.3 suggests, both approaches to the stepwise selection resulted in two different models. It is also important to note that **Model 1.2** is nested in **Model 1.1**. While the percentage of variance explained by **Model 1.1** is about 2% higher than that of **Model 1.2**, the latter's AIC value is smaller. In order to select the one that best fits the data besides looking at these measures, a test for nested models was applied, with a respective p-value of 0.06559. This leads to the conclusion that there is no significant difference between both models, and, therefore, the most parsimonious should be chosen: **Model 1.2**.

Table 3.4.3: Parameter estimates and their standard errors, observed values of the t-statistic and p-values for Model 1.2.

	Estimate	Standard Error	Observed t-statistic	p-value
Intercept	31.23635	2.05372	15.21000	< 0.00001
Psychological wellbeing (ASSET) (Ref: Very good)				
Good health levels	-2.24484	0.83753	-2.68000	0.00787
Average health levels	-3.46645	0.75245	-4.60700	0.00001
Low health levels	-5.62134	1.14954	-4.89000	0.00000
Very low health levels	-6.23638	1.37311	-4.54200	0.00001
Interests/hobbies (Ref: No)				
Yes	2.09814	0.80915	2.59300	0.01011

Doesn't know	-5.69255	3.41638	-1.66600	0.09698
Job security (ASSET) (Ref: Very low)				
Low levels of the stressor	-4.64156	1.33832	-3.46800	0.00062
Average levels of the stressor	-0.12203	0.90526	-0.13500	0.89289
High levels of the stressor	0.70107	1.25332	0.55900	0.57644
Very high levels of the stressor	-2.75424	1.54929	-1.77800	0.07673
Medication for chronic anxiety (Ref: No)				
Yes	-2.99778	0.89894	-3.33500	0.00099
Doesn't apply	-0.58589	0.46856	-1.25000	0.21239
Doesn't know	-2.80705	1.78424	-1.57300	0.11700
Presenteeism	0.05095	0.01694	3.00700	0.00293
Overload (ASSET) (Ref: Very low)				
Low levels of the stressor	-0.93544	1.05215	-0.88900	0.37486
Average levels of the stressor	-2.44519	0.82767	-2.95400	0.00345
High levels of the stressor	-3.27139	1.12084	-2.91900	0.00385
Very high levels of the stressor	-3.51630	1.18634	-2.96400	0.00335

Besides the significant differences between groups already accounted for regarding Table 3.4.1, the p-values of Table 3.4.4 lead to the reinforcement of the significant relationship between resilience and presenteeism, and show there are also significant differences between the subjects who take medication for chronic anxiety or who do not know and those who do not take medication for this condition.

Given these results and considering **Model 1.2** the final regression model to explain resilience for this dataset, the assumption validation and discordant observations analysis must be conducted, to ensure the model's reliability.

Assumption validation

- Normality of errors

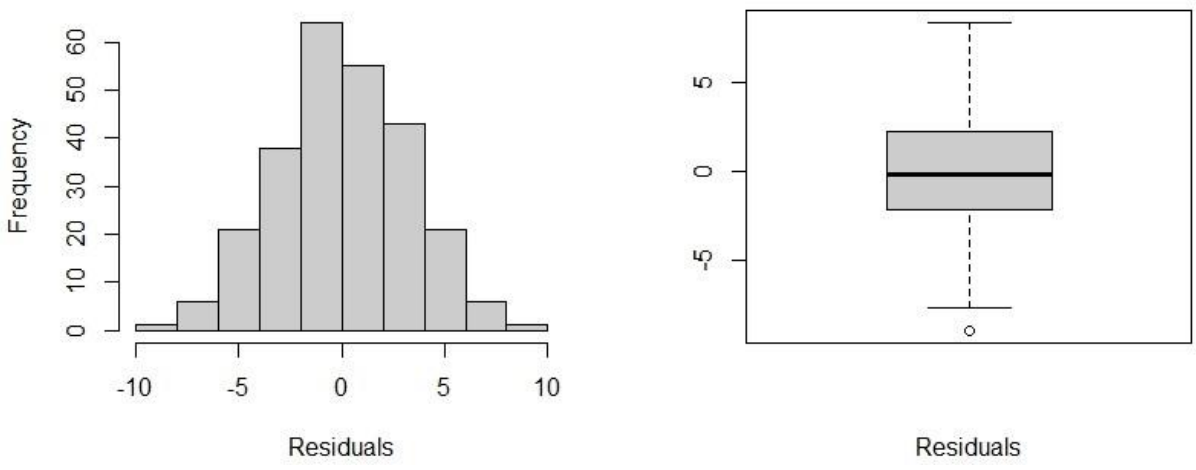

Figure 3.4.1: Histogram (left) and boxplot (right) of Model 1.2's residuals.

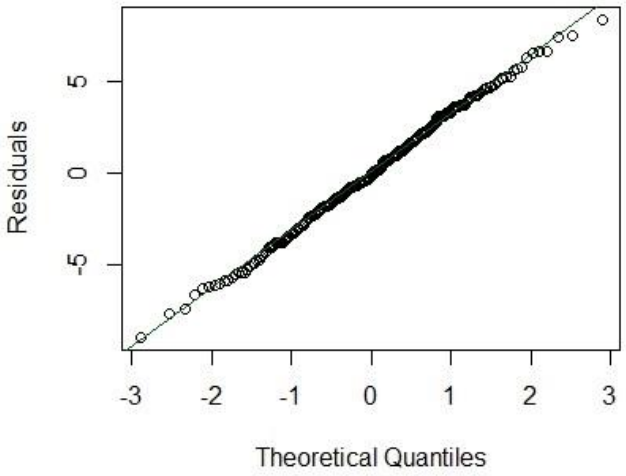

Figure 3.4.2: Quantile-quantile plot of Model 1.2's residuals.

The residuals histogram in Figure 3.4.1 appears to have a bell-shape, similar to a Normal distribution, and the boxplot to its right is approximately symmetrical around zero, apart from one outlier candidate. These results are verified by the quantile-quantile plot (Figure 3.4.2), which shows an approximately linear relationship between the model's theoretical quantiles and its residuals. From these graphic representations, the normality of residuals can be assumed.

- Homoscedasticity of errors

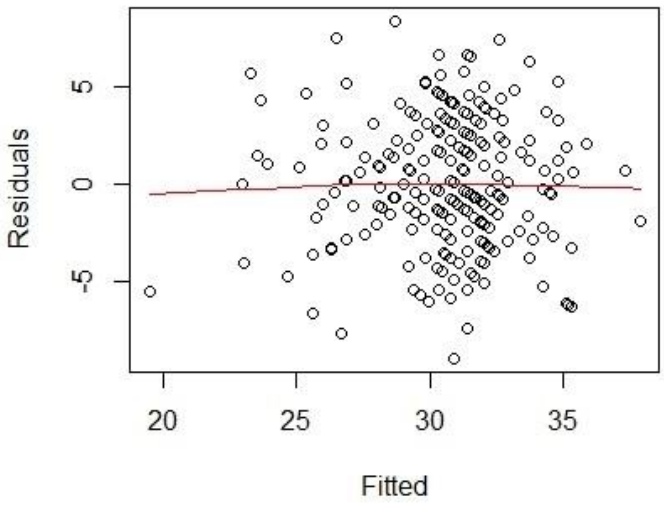

Figure 3.4.3: Scatter plot of predicted values *vs* residuals of Model 1.2.

Figure 3.4.3 suggests that the model's residuals are uniformly distributed around zero, and no distinct shape can be detected. Accordingly, the homoscedasticity of this model's residuals can be assumed.

- Linear relationship between the dependent variable and the independent variables

  As most of the variables in **Model 1.2** are categorical with more than two groups, there is no way to assess if a linear relationship exists between them and resilience. For the only numerical variable in this model, presenteeism, a rank correlation coefficient has already been computed in Section **3.3.1**, which indicates there is a significant rank correlation between resilience and presenteeism, despite weak.

- Independence of errors

  Based on Figure B.1, there seems to be no indication that the values of CDRISC-10 are influenced by the order in which the questionnaires were applied. Therefore, the independence of errors can be assumed.

- Absence of multicollinearity

Table 3.4.4.: VIFs of the independent variables of Model 1.2.

Independent variable	VIF
Psychological wellbeing (ASSET)	
Very low health levels	1.65059
Low health levels	1.91070
Average health levels	3.10758
Good health levels	2.37025
Interests/hobbies	
Yes	1.14984
Doesn't know	1.05918
Job security	
Low levels of the stressor	1.56802
Average levels of the stressor	2.95809
High levels of the stressor	2.02082
Very high levels of the stressor	1.48889
Medication for chronic anxiety	
Yes	1.29497
Doesn't know	1.14199
Doesn't apply	1.25755
Presenteeism	1.20264
Overload	
Low levels of the stressor	1.94419
Average levels of the stressor	3.36010
High levels of the stressor	2.01320
Very high levels of the stressor	1.92329

As no VIF presented in Table 3.4.6 is higher than 4, the absence of multicollinearity assumption is validated for this model.

Analysis of discordant observations

- Outliers

  The outlier test described in Section **2.7** presents a corrected p-value of approximately 1, which leads to the conclusion that the subject with the highest absolute residual is not an outlier, and therefore, none of the remaining subjects are either.

- Influential observations

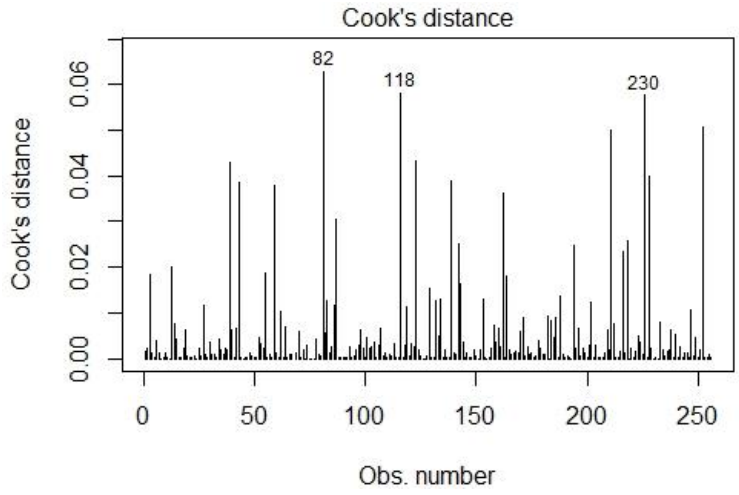

Figure 3.4.4: Cook's distances for Model 1.2.

As verified in Figure 3.4.4, the largest Cook's distances take very small values, very inferior to the critical value of 0.5. This means these individuals are not considered to be influential observations.

3.4.1 Multiple Linear Regression model with interactions

In order to check if **Model 1.2** could be improved by including plausible interactions between appropriate pairs of variables, the Fisher's exact test and Kruskal-Wallis tests were used. By applying Fisher's exact test, there was statistical evidence to assume that the following pairs of categorical dependent variables are associated:

- Psychological wellbeing and Job security;

- Psychological wellbeing and Medication for chronic anxiety;

- Psychological wellbeing and Overload;

- Job security and Overload.

By applying the Kruskal-Wallis test, it was also possible to detect statistically significant differences of Presenteeism between the groups of Psychological wellbeing, Overload and Medication for chronic anxiety.

Given these results, the respective interaction terms were added to the regression model, but none of them were significant. Therefore, **Model 1.2** stands as the final model.

3.5 Multiple Factor Analysis of Mixed Data

In this analysis, the $p = 20$ chosen variables in Table 3.3.1.1 and Table 3.3.2.1 were divided in $G = 2$ conceptual groups: "Mental and Physical Health" and "Work". Each group contains both numeric and categorical variables (Table 3.5.1).

Table 3.5.1: Groups of variables and their types for the implementation of MFA of mixed data.

Group	Variable	Type
Mental and Physical Health	Interests/hobbies	Categorical
	Psychological distress (MHI-5)	Categorical
	Psychological wellbeing (ASSET)	Categorical
	Physical health (ASSET)	Categorical
	Subjective Happiness Scale	Numerical
	Chronic anxiety	Categorical
	Depression	Categorical
	Medication for chronic anxiety	Categorical
	Medication for anxiety in the previous 2 weeks	Categorical
Work	Work relationships (ASSET)	Categorical
	Overload (ASSET)	Categorical
	Job security (ASSET)	Categorical
	Control (ASSET)	Categorical
	Resources and communication (ASSET)	Categorical
	Aspects of the job (ASSET)	Categorical
	Perceived commitment of employee to organization (ASSET)	Categorical
	Job Satisfaction Scale	Numerical
	Presenteeism	Numerical
	Number of missed days	Numerical
	Productivity	Categorical

Table 3.5.2: Eigenvalues, percentage of explained variance and cumulative percentage of explained variance for the first ten principal components of the MFA for mixed data.

Principal component	Eigenvalue	Percentage (%) of explained variance	Cumulative percentage (%) of explained variance
1	1.46041	10.72323	10.72323
2	0.80996	5.94725	16.67049
3	0.64102	4.70675	21.37724
4	0.52029	3.82027	25.19751
5	0.50774	3.72813	28.92564
6	0.46558	3.41856	32.34420
7	0.40573	2.97909	35.32329
8	0.39535	2.90292	38.22621
9	0.38440	2.82251	41.04873
10	0.36919	2.71084	43.75957

From Table 4.5.2, it is clear that, even with as many as 10 principal components, the cumulative percentage of explained variance is quite small (less than 50%). Given these results and based on the percentage of explained variance for each of the principal components, the first three are kept. These explain a little over 20% of the total variance of the data, and the percentage of explained variance of the fourth principal component and so on decrease substantially.

Table 3.5.3: Eigenvalues of each variable group for the first three principal components.

Principal component	Mental and Physical Health	Work
1	3.68216	4.27918
2	2.00449	2.64155
3	1.57965	2.11088

The "Work" group is the one that contributes the most to each of the first three principal components. (Table 3.5.3).

Table 3.5.4: Percentage of explained variance of each variable for the first three principal components.

Variable type	Variable	Principal component 1	Principal component 2	Principal component 3
Categorical	Psychological distress (MHI-5)	8.56112	1.67159	1.57637
	Medication for anxiety in the previous 2 weeks	3.54911	4.06491	0.72535
	Chronic anxiety	7.28128	3.40268	20.14789
	Medication for chronic anxiety	6.40351	4.44828	13.86084
	Depression	4.64680	0.66137	5.86413
	Interests/hobbies	0.20698	1.42437	3.45237
	Productivity	4.83739	1.20327	6.40720
	Physical health (ASSET)	6.57012	6.37538	0.74910
	Psychological wellbeing (ASSET)	10.40132	6.40275	15.20328
	Work relationships (ASSET)	6.10777	14.16784	5.45193
	Overload (ASSET)	3.07527	10.46526	7.68237
	Job security (ASSET)	2.99721	7.17600	0.72229
	Control (ASSET)	6.58585	11.30818	1.75068
	Resources and communication (ASSET)	5.89613	11.11224	3.77860
	Aspects of the job (ASSET)	4.40198	9.29283	3.72943
	Perceived commitment of employee to organization (ASSET)	4.39899	3.69290	0.52823
Numerical	Subjective Happiness Scale	3.55444	0.98449	1.86606
	Job Satisfaction Scale	6.92132	1.52795	0.44343
	Presenteeism	2.97287	0.61469	4.91855
	Number of missed days	0.63054	0.00302	1.14190

According to Table 3.5.4, psychological wellbeing, psychological distress and chronic anxiety explain more than a fourth of the variance of the first principal component. For the second principal component, work relationships, control, resources and communication and overload explain over 45% of its

variance. Finally, chronic anxiety, psychological wellbeing and medication for chronic anxiety explain almost half of the third principal component's variance.

Table 3.5.5: Squared loadings of each variable for the first three principal components.

Variable	Principal component 1	Principal component 2	Principal component 3
Subjective Happiness Scale	0.19114	0.02936	0.04405
Job Satisfaction Scale	0.43254	0.05296	0.01216
Presenteeism	0.18578	0.02131	0.13492
Number of missed days	0.03940	0.00010	0.03132
Psychological distress (MHI-5)	0.46037	0.04985	0.03721
Chronic anxiety	0.39155	0.10148	0.47556
Medication for chronic anxiety	0.34435	0.13267	0.32716
Anxiety medication in the previous 2 weeks	0.19085	0.12123	0.01712
Depression	0.24988	0.01972	0.13841
Interests/hobbies	0.01113	0.04248	0.08149
Productivity	0.30230	0.04171	0.17575
Psychological wellbeing (ASSET)	0.55933	0.19096	0.35885
Physical health (ASSET)	0.35331	0.19014	0.01768
Work relationships (ASSET)	0.38170	0.49105	0.14955
Overload (ASSET)	0.19218	0.36272	0.21073
Job security (ASSET)	0.18731	0.24872	0.01981
Control (ASSET)	0.41157	0.39194	0.04802
Resources and communication (ASSET)	0.36847	0.38515	0.10365
Aspects of the job (ASSET)	0.27509	0.32209	0.10230
Perceived commitment of employee to organization (ASSET)	0.27491	0.12799	0.01449

The results on Table 3.5.5 suggest that almost all variables contribute the most to the first principal component, except for work relationships, resources and communication, overload and job security, who present the highest contributions to the second principal component, and chronic anxiety and productivity, who contribute the most to the third principal component.

Graphical outputs

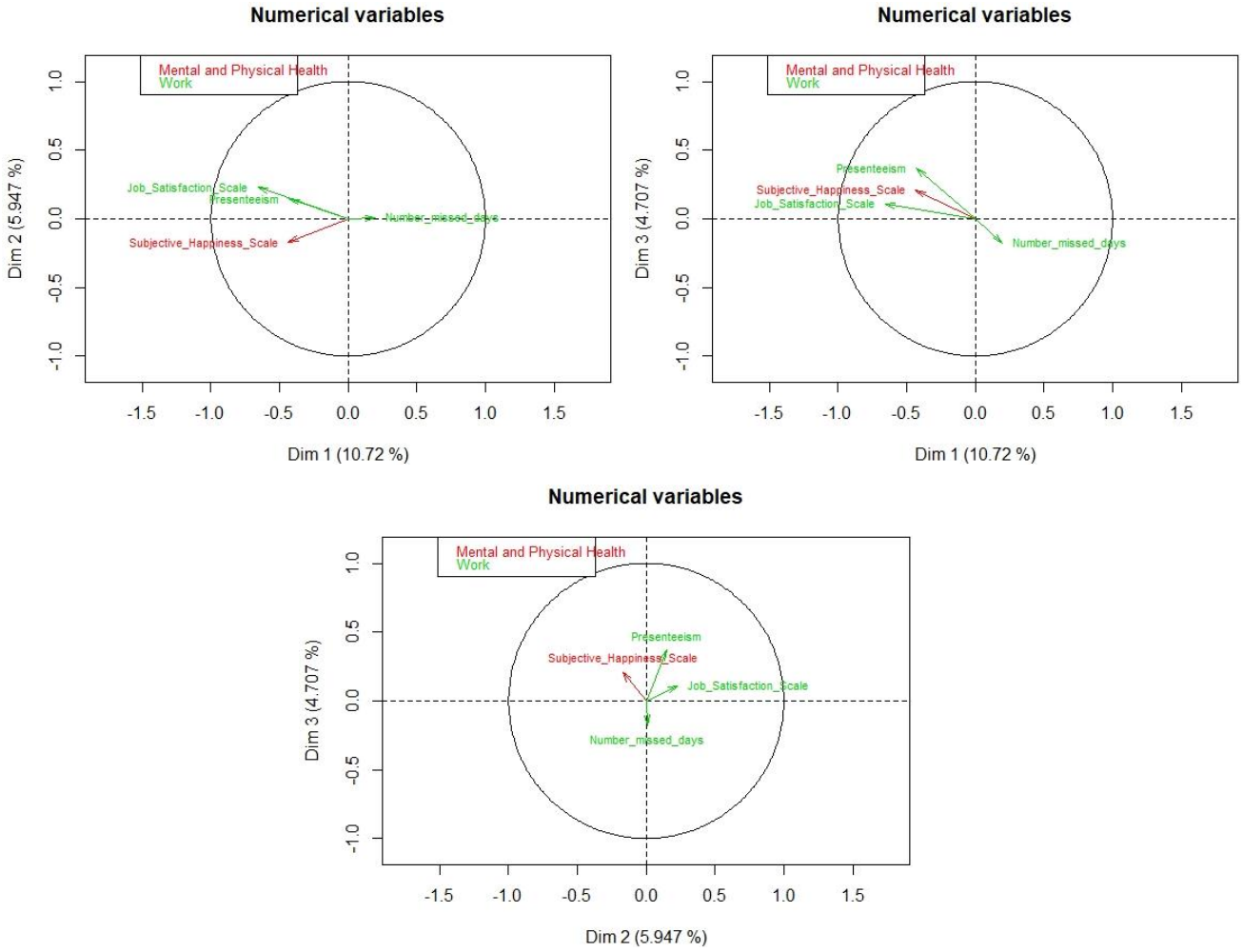

Figure 3.5.1: Correlation circles of the numerical variables for the combinations of the first three principal components.

Figure 3.5.1 contains the correlation circles of the 4 numerical variables, colored according to their group membership. The coordinates of the variables on this map represent their correlations with each principal component. The variable “Number of missed days” has a positive correlation with the first principal component, a negative correlation with the third principal component and no correlation with the second principal component. The remaining variables are negatively correlated with the first principal component, while for the second component this is only true for the Subjective Happiness, and positively correlated with the third principal component.

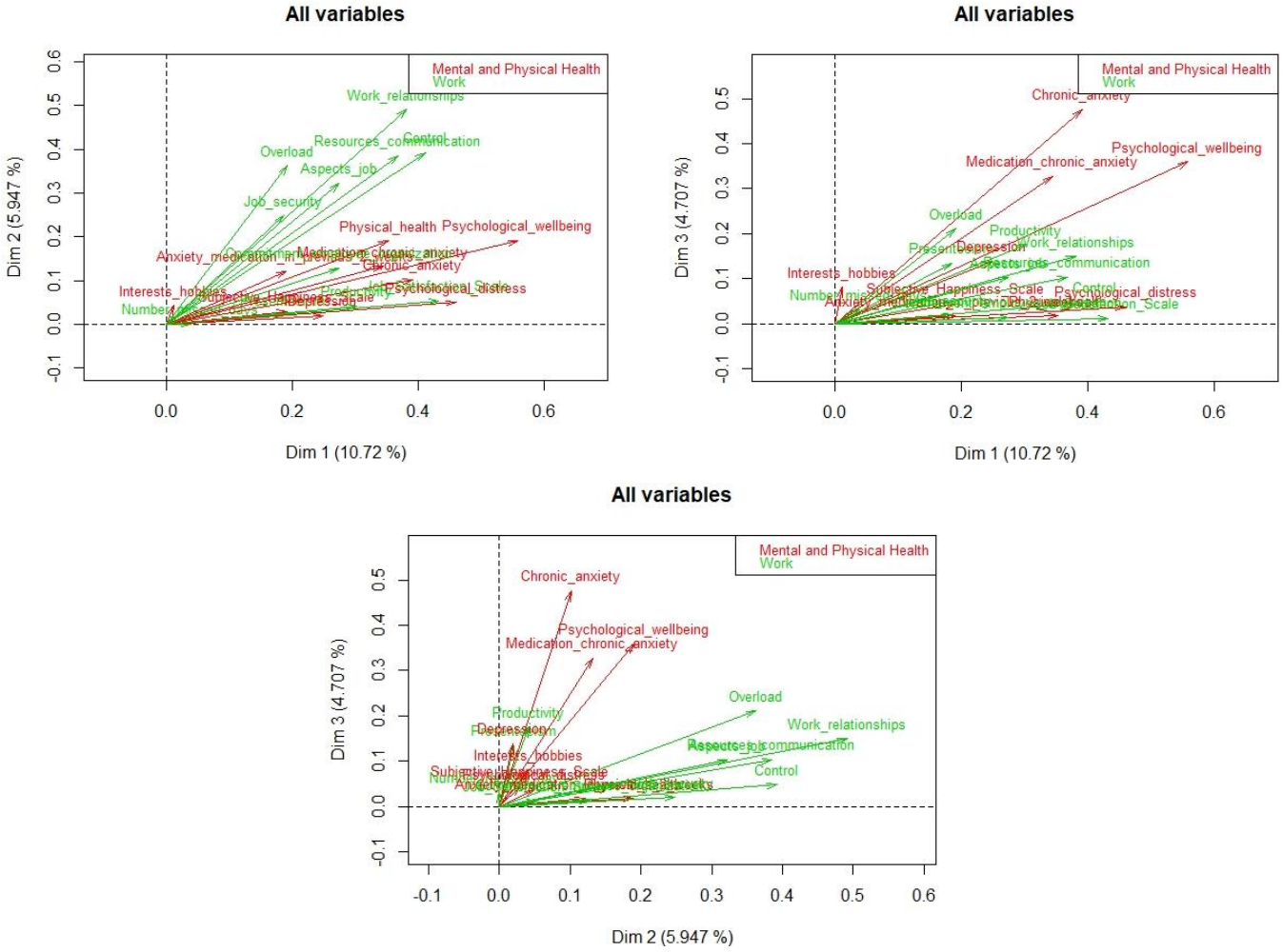

Figure 3.5.2: Contribution of each variable for the combinations of the first three principal components.

The maps in Figure 3.5.2 show that most of the variables in the "Work" group present high contributions to the first and second principal components and very little to the third, while the majority of the variables in the "Mental and Physical Health" group contribute strongly to the first and third principal components.

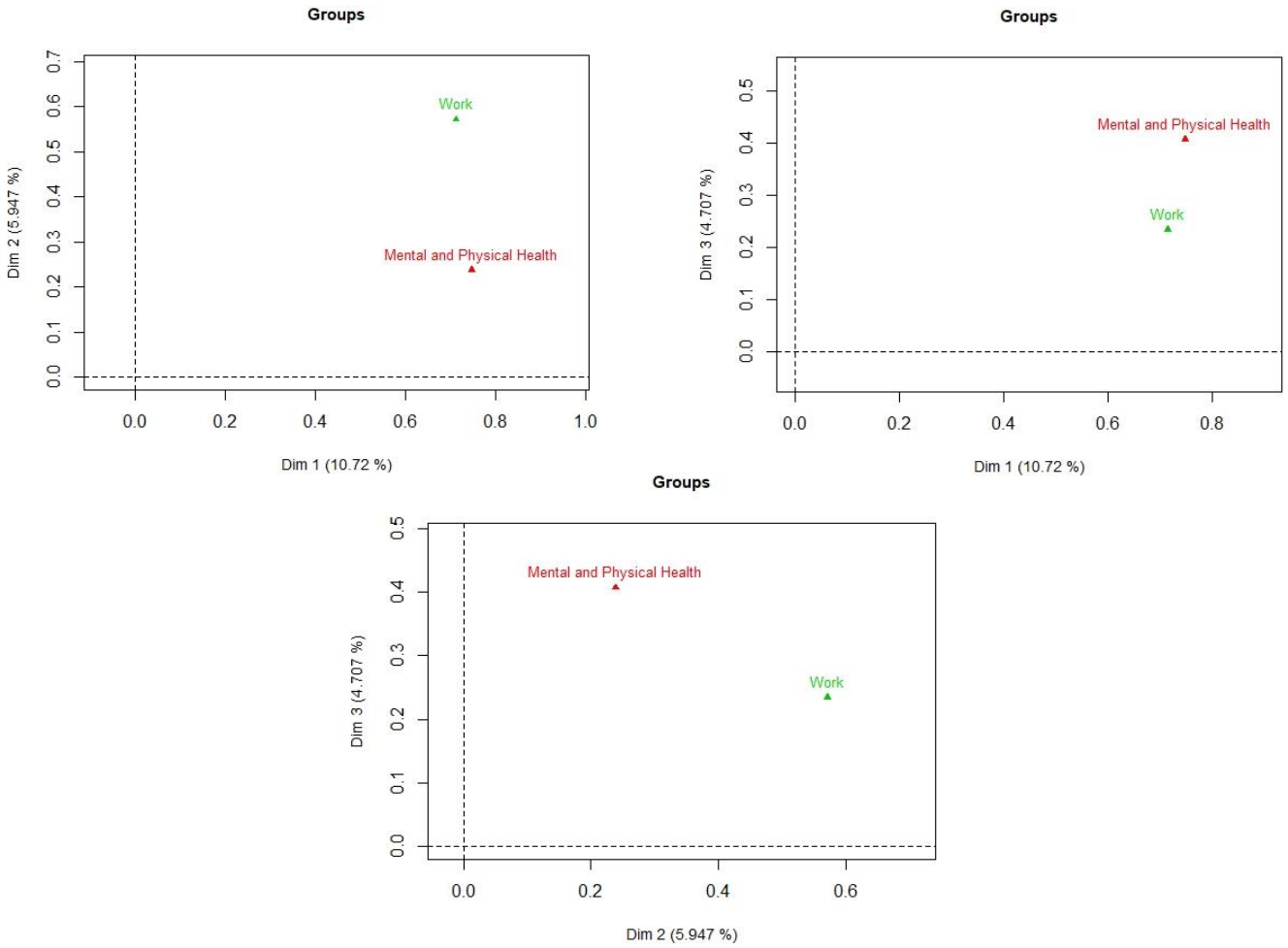

Figure 3.5.3: Contribution of each group for the combinations of the first three principal components.

The graphs of Figure 3.5.3 show that, regarding the first principal component, the impact of both groups is almost identical. The "Work" group contributes the most for the second principal component, while for the third the "Mental and Physical Health" group bears the highest contribution, although the difference is quite small.

Discussion

Summary of key findings

The main objective of this study was to identify the most relevant psychosocial and biological predictors for resilience in workers. To achieve this goal, a set of well documented and peer approved statistical techniques were applied. The results indicate that both internal and environmental factors may be linked to resilience and influence its amount.

From the initial variable selection, conducted through Mann-Whitney and Kruskal-Wallis tests and Spearman's rank correlation coefficient, most of ASSET's subscales (work relationships, overload, job security, control, resources and communication, aspects of the job, perceived commitment of employee to organization, physical health and psychological wellbeing), the MHI-5 scale (psychological distress), the Job Satisfaction Scale, the Subjective Happiness Scale, the Presenteeism scale, having interests/hobbies, having had depression and chronic anxiety, having taken medication for chronic anxiety and anxiety medication in the previous 2 weeks, perceived work productivity and the number of missed work days are all plausible resilience predictors.

Through a stepwise model selection, that took place to determine the most parsimonious regression model containing the previously selected variables, it was possible to assess that the most important variables for resilience prediction are the psychological wellbeing, job security and overload ASSET subscales, having interests/hobbies, having taken medication for chronic anxiety and the percentage of work performance loss. This final regression model explains about 35% of resilience's variability.

No significant results were found to associate resilience and the physiological or biochemical measurements, except for some genetic markers that were not explored in the current project, which will be developed in the context of a paper that is still in course.

In order to study the existence of a structure regarding the pre-selected variables that presented a significant relationship with resilience, a multiple factor analysis of mixed data was conducted. This method allows the presence of conceptual groups of variables, which can accommodate both numerical and categorical variables. This is important due to the large quantity of categorical variables usually found in surveys. The first three principal components explain approximately 20% of the data. Regarding the first principal component, the impact of both groups is almost identical, although the numerical variable "Number of missed days" of the "Work" group has a negative correlation with this component, unlike the remaining numerical variables ("Subjective Happiness Scale", "Job Satisfaction Scale" and "Presenteeism"). The same goes for the third principal component, revealing that among the numerical variables, the number of days a person is absent from work has a negative effect on resilience. The "Work" group contributes the most to the second principal component, while the "Mental and Physical Health" group makes the largest contribution to the third principal component, although the difference is quite small.

Discussing the results

The presented results build on existing evidence that mental health conditions like depression and anxiety may have a negative impact in resilience levels, while positive emotions seem to be resilience-promoting, according to Tugade and Fredrickson (2004) and Bonanno et al. (2002), respectively. However, they contradict the claims of Sparks et al. (1997) regarding variables such as sex, education level and income level, and certain conclusions of Bonanno et al. (2007) regarding age and social support, whose associations with resilience are not statistically significant in this study.

Due to the comprehensive questionnaire applied, this study also provides a new insight into the relationships between resilience and stress, work, happiness and mental health-related scales. These can be a viable addition to future surveys done on this subject, as a way of condensing important indicators and using validated questions to increase plausibility.

Most of the reported inconsistent results may be attributed to differences in resilience questionnaires, sample sizes and type of human subjects. A standardization of these is critical to discern the truly important and decisive factors underlying the resilience phenomenon.

Limitations

The generalization of the results obtained in this study is limited, mainly due to the data source and the small sample size. Besides the fact that the analyzed subjects represent less than 20% of the target population, the institution from where the data were collected is a very specific type of workplace, which makes it unlikely that it can represent other kinds of workplaces, at least regarding the study of resilience. Also, as the methods applied in this study were only able to explain a small percentage of resilience and its predictors' structure, it seems like the types of variables included in the questionnaire only represent a part of what influences workers' resilience.

The reliability of this data may be impacted by the non-randomness of the sample, since most of the statistical methods assume the use of a random sample. This assumption was considered to be true for this analysis, to allow the application of techniques that are generally accepted by the statistical community.

The methodological choices were constrained by the original study design. As a cross-sectional study, it is only capable of presenting a picture of a population's characteristics at a certain time. The fact that all the scales contained in the questionnaire are already validated also reduced the amount of available statistical methods to utilize as far as this type of study goes.

Despite these limitations, the presented results can be considered valid within the context of the study itself and for the purpose of answering its research questions. By the conducted analysis, it was possible to discover associations and predictive qualities between resilience and several scales, namely the ASSET subscales, which had not been done in other studies. This is a clear indication that much of what influences employees' resilience is related to their work.

The future study of resilience

Based on this study's results and specified limitations, further research should take into account a larger sample, if possible obtained randomly and from diversified workplaces, as well as a larger spectrum of variables, in order to not only ensure the results' generalization and reliability to a larger population of workers, but also to evaluate dimensions not included in the present study. The research on the resilience phenomenon is still scarce, so further investigation is needed to generate stronger evidence regarding its possible predictors. This will not only facilitate the understanding of workers' habits, performance and behaviors, but will also allow companies and health professionals to better help unmotivated and unproductive employees.

References

Abdi, H., Williams, L. J., Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2): 149-179.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57: 289–300.

Black J. K., Balanos G. M., Whittaker A. C. (2017). Resilience, work engagement and stress reactivity in a middle-aged manual worker population. *Int. J. Psychophysiol*. 116: 9–15. doi:10.1016/j.ijpsycho.2017.02.013

Bonanno G. A., Wortman, C. B., Lehman, D. R., Tweed, R. G., Haring, M. et al. (2002). *J Pers Soc Psychol*. 83(5): 1150-1164.

Bonanno, G. A., Galea, S., Bucciarelli, A., Vlahov, D. (2007). What Predicts Psychological Resilience After Disaster? The Role of Demographics, Resources, and Life Stress. *Journal of Consulting and Clinical Psychology*. 75(5): 671-682.

Chavent, M., Kuentz-Simonet, V., Labenne, A., Saracco, J. (2014). Multivariate analysis of mixed data: The R Package PCAmixdata, https://arxiv.org/abs/1411.4911

Colman, A. (2019). A Dictionary of Psychology. Oxford: Oxford University Press.

Connor, K. M., Davidson, J. R., Lee, L. C. (2003). Spirituality, resilience, and anger in survivors of violent trauma: a community survey. *J Trauma Stress*. 16: 487–94. doi:10.1023/A:1025762512279

Connor, K. M., Davidson, J. R. T. (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depression and Anxiety*, 18(2): 76–82.

Conover, W. J., Iman, R. L. (1979). On multiple-comparisons procedures. Technical Report LA-7677-MS, Los Alamos Scientific Laboratory.

Cooper, C. L., Sloan, S. J., & Williams, S. (1988). Occupational Stress Indicator. Nefer-Nelson

Corina D., Adriana B. (2013). Impact of work related trauma on acute stress response in train drivers. *Procedia Soc. Behav. Sci.* 84: 190–195. doi:10.1016/j.sbspro.2013.06.533

Dalgard, O. S. (1996). Community health profile as tool for psychiatric prevention. In Promotion of mental health, D. R. Trent & C. Reed (Eds.). Aldershot, Avebury.

Dalgard, O. S., Dowrick, C., Lehtinen, V., Vazquez-Barquero, J. L., Casey, P., Wilkinson, G., Dunn, G. (2006). Negative life events, social support and gender difference in depression. *Social Psychiatry and Psychiatric Epidemiology*, 41(6): 444–451.

Escofier, B., Pagès, J. (1994). Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis*, 18(1): 121-140.

Everitt, B. (1998). The Cambridge Dictionary of Statistics. Cambridge, UK New York: Cambridge University Press. ISBN 0-521-59346-8.

Ewing, J. A. (1984). Detecting alcoholism: the CAGE questionnaire. *Jama*, 252(14): 1905–1907.

Faria Anjos, J., Heitor dos Santos, M.J., Ribeiro, M.T., Moreira, S. (2019). Connor-Davidson. Resilience Scale: validation study in a Portuguese sample. *BMJ Open*, 9(6). doi:10.1136/bmjopen-2018-026836.

Fayombo, G. (2010). The Relationship between Personality Traits and Psychological Resilience among the Caribbean Adolescents. *International Journal of Psychological Studies*. 2. doi:10.5539/ijps.v2n2p105.

Fowler, J., Cohen, L., Jarvis, P. (2009). Practical Statistics for Field Biology. p. 132

George Mason University's Resilience Model. (n.d.). https://wellbeing.gmu.edu/resources/george-mason-university-s-resilience-model

Hauke, J., Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2): 87.

Heitor dos Santos, M.J., Moreira, S., Carreiras, J., Cooper, C., Smeed, M., Reis, M.F., Pereira Miguel, J. (2018). Portuguese version of a stress and wellbeing evaluation tool (ASSET) at the workplace: Validation of the psychometric properties. *BMJ Open*, 8. doi:10.1136/bmjopen-2017-018401.

Jacelon, C. S. (1997) The trait and process of resilience. *Journal of Advanced Nursing*, 25: 123-129. doi:10.1046/j.1365-2648.1997.1997025123

Jackson, D., Firtko, A., Edenborough, M. (2007). Personal resilience as a strategy for surviving and thriving in the face of workplace adversity: a literature review. *Journal of Advanced Nursing*, 60(1).

Kessler, R. C., Barber, C., Beck, A., Berglund, P., Cleary, P. D., McKenas, D., Ustun, T. B. (2003). The world health organization health and work performance questionnaire (HPQ). *Journal of Occupational and Environmental Medicine*, 45(2): 156–174.

Kessler, R. C., Ames, M., Hymel, P. A., Loeppke, R., McKenas, D. K., Richling, D. E., … Ustun, T. B. (2004). Using the World Health Organization Health and Work Performance Questionnaire (HPQ) to evaluate the indirect workplace costs of illness. *Journal of Occupational and Environmental Medicine*, 46(6): S23–S37.

Kruskal, W.H., Wallis, W.A. (1952). "Use of ranks in one-criterion variance analysis". *Journal of the American Statistical Association*. 47(260): 583–621. doi:10.1080/01621459.1952.10483441

Levene, H. (1960). "Robust tests for equality of variances". In Ingram Olkin; Harold Hotelling; et al. (eds.). Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press. 278–292.

Liu, H., Zhang, C., Ji, Y., Yang, L. (2018). Biological and Psychological Perspectives of Resilience: Is It Possible to Improve Stress Resistance? *Front Hum Neurosci*. 12(326). doi:10.3389/fnhum.2018.00326

Lyubomirsky, S., Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, 46(2): 137–155.

Mann, H. B., Whitney, D. R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". *Annals of Mathematical Statistics*. 18(1): 50–60. doi:10.1214/aoms/1177730491

Mehta, C.R., Patel, N.R. (1983). "A Network Algorithm for Performing Fisher's Exact Test in r Xc Contingency Tables". *Journal of the American Statistical Association*. 78(382): 427–434. doi:10.2307/2288652

Merriam-Webster: America's most-trusted online dictionary. (2019). https://www.merriam-webster.com.

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, 89(3): 570–574. doi:10.1037/0033-2909.89.3.570

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Razali, N., Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics*. 2(1): 21–33.

Rees, C. S., Breen, L. J., Cusack, L., Hegney, D. (2015). Understanding individual resilience in the workplace: the international collaboration of workforce resilience model. *Frontiers in Psychology*, 6(73). doi:10.3389/fpsyg.2015.00073

Ribeiro, J. L. P. (2001). Mental health inventory: Um estudo de adaptação à população portuguesa. *Psicologia, Saúde E Doenças*, 2(1): 77–99.

Rutter, M. (2006). Implications of Resilience Concepts for Scientific Understanding. *Annals of the New York Academy of Sciences*, 1094: 1-12.

Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4): 591–611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709

Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods*, 11(2): 387–407. doi:10.1177/1094428106292901

Sparks, K., Cooper, C., Fried, Y., Shirom, A. (1997). The effects of hours of work on health: A meta-analytic review. *Journal of Occupational and Organizational Psychology*, 70: 391–408.

Tugade, M. M., Fredrickson, B. L. (2004). Resilient individuals use positive emotions to bounce back from negative emotional experiences. *Journal of Personality and Social Psychology*, 86, 320 – 333.

Veit, C. T., Ware, J. E. (1983). The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*, 51(5): 730–742.

Weisstein, E. W. (n.d.) "Fisher's Exact Test". MathWorld - A Wolfram Web Resource. http://mathworld.wolfram.com/FishersExactTest.html

Weiner, I. B., Schmitt, N. W., Highhouse, S. (2012). Handbook of Psychology. Wiley

World Health Organization. (1994). Global strategy on occupational health for all: The way to health at work. https://www.who.int/occupational_health/publications/globstrategy/en/index2.html

Appendices

A

List of variables

Table A.1: List of all the final variables in the dataset.

Variable	Type	Labels/Units
Study ID	Numerical	
Sociodemographic		
Sex	Categorical	Female, Male
Age	Numerical	
Education	Categorical	Primary education, Secondary education, College education, Postgraduate education, Other
Marital status	Categorical	Single, Single in a non-marital partnership, Married/In a civil union, Divorced/Separated, Widowed
Number of children	Categorical	0, 1, 2, 3+
Living alone	Categorical	Yes, No
Plans to have children	Categorical	I'm pregnant/going to be a father, Definitely yes, Probably yes, Probably not, Definitely not, Doesn't know
Work-related		
Functional Group	Categorical	White collars (level 1), White collars (level 2), Blue collars
Time until next promotion	Categorical	Within 1 year, 1-5 years, Over 5 years, Never, Doesn't know
Absenteeism	Numerical	
Number of missed days in the last 12 months	Numerical	
Monthly income	Categorical	486-986€, 987-1500€, 1501-2014€, 2015-2528€, 2529-3042€, 3043-3556€, +3557€
Productivity in the previous 3 months	Categorical	100%, 90-99%, 80-89%, 70-79%, <70%
Impact of the economic crisis in a scale of 0 to 10	Numerical	
Socialization with work colleagues	Categorical	Yes, No, Doesn't know
Lifestyle-related		

Implementation of a regular exercise regimen	Categorical	Always, Regularly, When possible, Occasionally, Rarely, Never
Smoker	Categorical	Yes, No, Doesn't know
Number of cigarettes per day	Categorical	1-5, 6-10, 11-20, 21-30, 31-40, 41+, Doesn't apply, Doesn't know
Smoking amount	Categorical	More than usual, The same amount, Less than usual, Doesn't apply, Doesn't know
Alcohol drinker	Categorical	Yes, No, Doesn't know
Number of alcoholic drinks per day	Categorical	1-5, 6-10, 11-20, 21-30, 31-40, 41+, Doesn't apply, Doesn't know
Alcohol amount	Categorical	More than usual, The same amount, Less than usual, Doesn't apply, Doesn't know
Coffee drinker	Categorical	Yes, No, Doesn't know
Number of coffees per day	Categorical	1-2, 3-4, 5-6, 7-8, 9+, Doesn't apply, Doesn't know
Coffee amount	Categorical	More than usual, The same amount, Less than usual, Doesn't apply, Doesn't know
Physical or mental disabilities	Categorical	Yes, No, Doesn't know
Time to relax	Categorical	Always, Frequently, When possible, Rarely
Interests/hobbies	Categorical	Yes, No, Doesn't know
Clinical		
Important diseases in the previous 6 months	Categorical	Yes, No
Global health state in the previous 3 months	Categorical	Good, Reasonable, Bad
Disturbing events in the previous 6 months	Categorical	Yes, No, Doesn't know
Depression	Categorical	Yes, No, Doesn't know
Chronic anxiety	Categorical	Yes, No, Doesn't know
Medication for depression	Categorical	Yes, No, Doesn't apply, Doesn't know
Medication for chronic anxiety	Categorical	Yes, No, Doesn't apply, Doesn't know
Number of missed days due to disease or accident	Categorical	0, 1, 2-5, 6+
Medication for high blood pressure in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for other cardiovascular diseases in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for joint pain in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for headaches in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for anxiety in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for depression in the previous 2 weeks	Categorical	Yes, No, Doesn't know

Medication for stomach issues in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for sleep in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for cholesterol in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Medication for diabetes in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Non-prescribed medication for pain in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Non-prescribed medication for stomach issues in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Non-prescribed medication for anxiety in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Non-prescribed medication for sleep in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Non-prescribed medication for fortifiers in the previous 2 weeks	Categorical	Yes, No, Doesn't know
Biochemical		
Hemoglobin	Numerical	g/dL
Red Blood Cell count (RBC)	Numerical	x1012/L
Hematocrits (HTC)	Numerical	L/L
Mean Corpuscular Volume (MCV)	Numerical	fl
Mean Corpuscular Hemoglobin (MCH)	Numerical	pg
Mean Corpuscular Hemoglobin Concentration (MCHC)	Numerical	g/dL
Red cell Distribution Width (RDW)	Numerical	%
GB	Numerical	x109/L
Neutrophils	Numerical	%
Lymphocytes	Numerical	%
Monocytes	Numerical	%
Eosinophils	Numerical	%
Basophils	Numerical	%
Platelets	Numerical	x106/L
Glucose	Numerical	mg/dL
Urea	Numerical	mg/dL
Creatinine	Numerical	mg/dL
Total Cholesterol	Numerical	mg/dL
High-density lipoprotein (HDL)	Numerical	mg/dL

Low-density lipoprotein (LDL)	Numerical	mg/dL
Triglycerides	Numerical	mg/dL
Aspartate Aminotransferase (AST)	Numerical	U/L
Alanine Aminotransferase (ALT)	Numerical	U/L
Alkaline Phosphatase	Numerical	U/L
Lactate Dehydrogenase	Numerical	U/L
Gamma-Glutamyl Transferase (GGT)	Numerical	U/L
C-Reactive Protein (CRP)	Numerical	mg/dL
Heterophils (HET)	Numerical	µU/L
DNA	Numerical	ng/µL
Body Mass Index (BMI)	Numerical	kg/m^2
Systolic blood pressure	Numerical	mmHg
Diastolic blood pressure	Numerical	mmHg
Heart rate	Numerical	
Cardiovascular Risk	Numerical	
Scales		
Work relationships (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Work-life balance (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Overload (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Job security (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Control (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Resources and communication (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Aspects of the job (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High

		levels of the stressor, Very high levels of the stressor
Pay and benefits (ASSET)	Categorical	Very low levels of the stressor, Low levels of the stressor, Average levels of the stressor, High levels of the stressor, Very high levels of the stressor
Perceived commitment of organization to employee (ASSET)	Categorical	Very low levels of commitment, Low levels of commitment, Average levels of commitment, High levels of commitment, Very high levels of commitment
Perceived commitment of employee to organization (ASSET)	Categorical	Very low levels of commitment, Low levels of commitment, Average levels of commitment, High levels of commitment, Very high levels of commitment
Physical health (ASSET)	Categorical	Very good health levels, Good health levels, Average health levels, Low health levels, Very low health levels
Psychological wellbeing (ASSET)	Categorical	Very good health levels, Good health levels, Average health levels, Low health levels, Very low health levels
Psychological distress (MHI-5)	Categorical	Yes, No
Job Satisfaction Scale	Numerical	
Presenteeism	Numerical	
OSLO-3	Categorical	Poor support, Moderate support, Strong support
Subjective Happiness Scale	Numerical	
CAGE scale	Categorical	Non-significant CAGE, Significant CAGE, Doesn't apply
CDRISC-25	Numerical	
CDRISC-10	Numerical	

B

Graphical validation of the independence of errors assumption

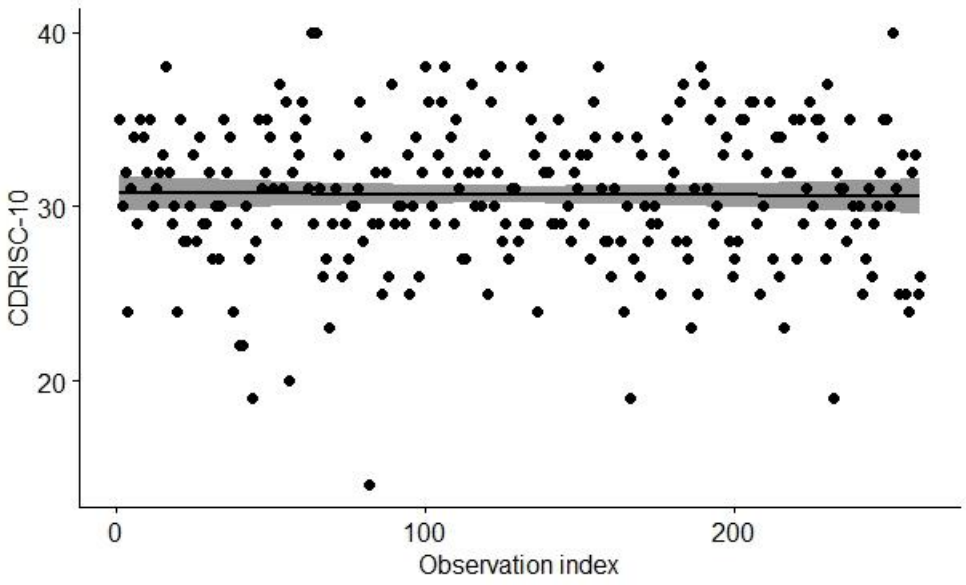

Figure B.1: Scatterplot of observation indices *vs* observed values of CDRISC-10.